JAEGWON KIM

PHILOSOPHY OF MIND

THIRD EDITION

VISIT…

PHILOSOPHY OF MIND

PHILOSOPHY OF MIND

THIRD EDITION

JAEGWON KIM

A Member of the Perseus Books Group

Westview Press was founded in 1975 in Boulder, Colorado, by notable publisher and intellectual Fred Praeger. Westview Press continues to publish scholarly titles and high-quality undergraduate- and graduate-level textbooks in core social science disciplines. With books developed, written, and edited with the needs of serious nonfiction readers, professors, and students in mind, Westview Press honors its long history of publishing books that matter.

Copyright © 2011 by Westview Press

Published by Westview Press,
A Member of the Perseus Books Group

All rights reserved. Printed in the United States of America. No part of this book may be reproduced in any manner whatsoever without written permission except in the case of brief quotations embodied in critical articles and reviews. For information, address Westview Press, 2465 Central Avenue, Boulder, CO 80301.

Find us on the World Wide Web at www.westviewpress.com.

Every effort has been made to secure required permissions for all text, images, maps, and other art reprinted in this volume.

Westview Press books are available at special discounts for bulk purchases in the United States by corporations, institutions, and other organizations. For more information, please contact the Special Markets Department at the Perseus Books Group, 2300 Chestnut Street, Suite 200, Philadelphia, PA 19103, or call (800) 810-4145, ext. 5000, or e-mail special.markets@perseusbooks.com.

Designed by Trish Wilkinson
Set in 10.5 point Minion Pro

Library of Congress Cataloging-in-Publication Data

Kim, Jaegwon.
Philosophy of mind / Jaegwon Kim.—3rd ed.
p. cm.
ISBN 978-0-8133-4458-4 (alk. paper)
1. Philosophy of mind. I. Title.
BD418.3.K54 2011
128'.2—dc22
E-book ISBN 978-0-8133-4520-8 2010040944

10 9 8 7 6 5 4 3 2 1

Contents

Preface *ix*

1. Introduction **1**

What Is Philosophy of Mind? 2
Metaphysical Preliminaries 5
Mind-Body Supervenience 8
Materialism and Physicalism 11
Varieties of Mental Phenomena 14
Is There a "Mark of the Mental"? 17
For Further Reading 28
Notes 28

2. Mind as Immaterial Substance: Descartes's Dualism **31**

Descartes's Interactionist Substance Dualism 32
Why Minds and Bodies Are Distinct: Some Arguments 35
Princess Elisabeth Against Descartes 46
The "Pairing Problem": Another Causal Argument 50
Immaterial Minds in Space? 54
Substance Dualism and Property Dualism 56
For Further Reading 58
Notes 58

3. Mind and Behavior: Behaviorism **61**

The Cartesian Theater and the "Beetle in the Box" 63
What Is Behavior? 66
Logical Behaviorism: A Positivist Argument 68
A Behavioral Translation of "Paul Has a Toothache" 70
Difficulties with Behavioral Definitions 71
Do Pains Entail Pain Behavior? 76
Ontological Behaviorism 78
The Real Relationship Between Pain and Pain Behavior 80
Behaviorism in Psychology 82
Why Behavior Matters to Mind 86
For Further Reading 87
Notes 88

4. Mind as the Brain: The Psychoneural Identity Theory **91**

Mind-Brain Correlations 91
Making Sense of Mind-Brain Correlations 93
The Argument from Simplicity 98
Explanatory Arguments for Psychoneural Identity 102
An Argument from Mental Causation 110
Against Psychoneural Identity Theory 114
Reductive and Nonreductive Physicalism 122
For Further Reading 125
Notes 126

5. Mind as a Computing Machine: Machine Functionalism **129**

Multiple Realizability and the Functional Conception of Mind 130
Functional Properties and Their Realizers: Definitions 134
Functionalism and Behaviorism 136
Turing Machines 139
Physical Realizers of Turing Machines 144

Machine Functionalism: Motivations and Claims 147
Machine Functionalism: Further Issues 151
Can Machines Think? The Turing Test 156
Computationalism and the "Chinese Room" 160
For Further Reading 165
Notes 165

6. Mind as a Causal System: Causal-Theoretical Functionalism 169

The Ramsey-Lewis Method 170
Choosing an Underlying Psychology 172
Functionalism as Physicalism: Psychological Reality 177
Objections and Difficulties 179
Roles Versus Realizers: The Status of Cognitive Science 186
For Further Reading 189
Notes 190

7. Mental Causation 193

Agency and Mental Causation 195
Mental Causation, Mental Realism, and Epiphenomenalism 197
Psychophysical Laws and "Anomalous Monism" 202
Is Anomalous Monism a Form of Epiphenomenalism? 207
Counterfactuals to the Rescue? 209
Physical Causal Closure and the "Exclusion Argument" 214
The "Supervenience Argument" and Epiphenomenalism 217
Further Issues: The Extrinsicness of Mental States 220
For Further Reading 223
Notes 224

8. Mental Content 227

Interpretation Theory 228
The Causal-Correlational Approach: Informational Semantics 235
Misrepresentation and the Teleological Approach 239

Narrow Content and Wide Content: Content Externalism 241
The Metaphysics of Wide Content States 248
Is Narrow Content Possible? 251
Two Problems for Content Externalism 254
For Further Reading 258
Notes 259

9. What Is Consciousness? 263

Some Views on Consciousness 264
Nagel and His Inscrutable Bats 267
Phenomenal Consciousness and Access Consciousness 271
Consciousness and Subjectivity 280
Does Consciousness Involve Higher-Order Perception or Thought? 283
Transparency of Experience and Qualia Representationalism 289
For Further Reading 295
Notes 296

10. Consciousness and the Mind-Body Problem 301

The "Explanatory Gap" and the "Hard Problem" 302
Does Consciousness Supervene on Physical Properties? 306
Closing the Explanatory Gap: Reduction and Reductive Explanation 311
Functional Analysis and Reductive Explanation 315
Consciousness and Brain Science 317
What Mary, the Super Vision Scientist, Didn't Know 323
The Limits of Physicalism 326
For Further Reading 333
Notes 334

References *339*
Index *355*

Preface

It has been five years since the appearance of the second edition. Philosophy of mind remains a vibrant, thriving field, and this is a good time to update and improve the book.

As in the earlier editions, we explore a range of issues in the philosophy of mind, with the mind-body problem as the main focus. The specific issues taken up, and our general approach, belong to what is now called the metaphysics of mind, but our discussion touches on issues in the epistemology and language of mind, and at various points the implications of our considerations for the status of the cognitive and behavioral sciences are explored. However, this is not a book on the philosophy of psychology or cognitive science; nor is it concerned with the "analysis" of psychological language or concepts. Its principal subject is the nature of mind, its relationship to our bodily nature, and its place in a world that is essentially material.

The main new feature of this edition is an expanded coverage of consciousness: The single chapter on consciousness has been replaced by two chapters, one on the nature of consciousness and the other on its philosophical and scientific status. This reflects the ongoing surge of activity in consciousness studies in both philosophy and the sciences. Arguably, consciousness is now the most actively debated topic in philosophy of mind, and the boom shows no sign of slowing down. Partly to make room for the additional chapter on consciousness, the last chapter of the second edition, on reduction and physicalism, has been dropped, with some of the material absorbed into the second of the two chapters on consciousness.

Most of the remaining chapters have been augmented with new material in various ways. And I have done what I could to improve the readability and clarity of the writing. But the guiding ideas of the previous editions remain the

same. In particular, the chapters are intended to be readable as independent essays. Cross-references are provided mainly to help the reader; they should not interrupt narrative flow or continuity. Like most contemporary philosophical works, this book is argument-oriented and presents a point of view. Although it has been written with readers new to the field as the primary audience, it is not a passive, dispassionate survey of the field; where I have my opinions, the reader will know where I stand. Interestingly but perhaps unsurprisingly, it has proved more difficult to write about topics on which I don't have settled views of my own. I have tried, though, to present fair and balanced pictures of alternative approaches and perspectives. I will be pleased if the book serves as a stimulus to the reader to engage with the problems of the mind and try to come to terms with them. That, after all, is what writing books like this is all about.

Chiwook Won, my graduate assistant at Brown, has given me invaluable help, with efficiency, intelligence, and good cheer. I am grateful to Karl Yambert, my former Westview editor, for support and encouragement. Finally, I want to thank the philosophers who responded to Karl's request for feedback on the second edition as a course text. Their candid and perceptive comments were most informative and helpful.

Providence, Rhode Island
September 2010

CHAPTER 1

Introduction

In coping with the myriad things and events that come our way at every moment of our waking life, we try to organize them into manageable chunks. We do this by sorting things into groups—categorizing them as "rocks," "trees," "fish," "birds," "bricks," "fires," "rains," and countless other kinds—and describing them in terms of their properties and features as "large" or "small," "tall" or "short," "red" or "yellow," "slow" or "swift," and so on. A distinction that we almost instinctively apply to just about everything is whether it is a *living* thing. (It might be a dead bird, but still we know it is the *kind* of thing that lives, unlike a rock or a celadon vase, which couldn't be "dead.") There are exceptions, of course, but it is unusual for us to know what something is without at the same time knowing, or having some ideas about, whether it is a living thing. Another example: When we know a person, we almost always know whether the person is male or female.

The same is true of the distinction between things, or creatures, with a "mind" and those without a mind. This, too, is one of the most basic contrasts we use in our thoughts about things in the world. Our attitudes toward creatures that are conscious and capable of experiencing sensations like pain and pleasure are importantly different from our attitudes toward things lacking such capacities, mere chunks of matter or insentient plants, as witness the controversies about vegetarianism and scientific experiments involving live animals. And we are apt to regard ourselves as occupying a special and distinctive place in the natural world on account of our particularly highly developed mental capacities and functions, such as the capacity for abstract thoughts, self-consciousness, artistic sensibilities, complex emotions, and a capacity for rational deliberation and action. Much as we admire the miracle of the flora and fauna, we do not think that every living thing has a mind or that we need a

psychological theory to understand the life cycles of elms and birches or the behavior and reproductive patterns of amoebas. Except those few of us with certain mystical inclinations, we do not think that members of the plant world are endowed with mentality, and we would exclude many members of the animal kingdom from the mental realm as well. We would not think that planarians and gnats have a mental life that is fit for serious psychological inquiry.

When we come to higher forms of animal life, such as cats, dogs, and chimpanzees, we find it entirely natural to grant them a fairly rich mental life. They are surely *conscious* in that they experience *sensations*, like pain, itch, and pleasure; they *perceive* their surroundings more or less the way we do and use the information so gained to guide their behavior. They also *remember* things—that is, store and use information about their surroundings—and *learn* from experience, and they certainly appear to have *feelings* and *emotions*, such as fear, frustration, and anxiety. We describe their psychological life using the expressions we normally use for fellow human beings: "Phoebe is feeling cramped inside the pet carrier and all that traffic noise has made her nervous. The poor thing is dying to be let out."

But are the animals, even the more intelligent ones like horses and dolphins, capable of complex social emotions like embarrassment and shame? Are they capable of forming intentions, engaging in deliberation and making decisions, or performing logical reasoning? When we go down the ladder of animal life to, say, oysters, crabs, and earthworms, we would think that their mental life is considerably impoverished in comparison with that of, say, a domestic cat. Surely these creatures have sensations, we think, for they react in appropriate ways to noxious stimuli, and they have sense organs through which they gain information about what goes on around them and adjust and modify their behavior accordingly. But do they have minds? Are they conscious? Do they have mentality? What is it to have a mind, or mentality?

What Is Philosophy of Mind?

Philosophy of mind, like any other field of inquiry, is defined by a group of problems. As we expect, the problems that constitute this field concern mentality and mental properties. What are some of these problems? And how do they differ from the scientific problems about mentality and mental properties, those that psychologists, cognitive scientists, and neuroscientists investigate in their research?

There is, first of all, the problem of answering the question raised earlier: What is it to be a creature with a mind? Before we can fruitfully consider ques-

tions like whether inorganic electromechanical devices (for example, computers and robots) can have a mind, or whether speechless animals are capable of having thoughts, we need a reasonably clear idea about what mentality is and what having a thought consists in. What conditions must a creature or system meet if we are to attribute to it a "mind" or "mentality"? We commonly distinguish between mental phenomena, like thoughts and sensory experiences, and those that are not mental, like digestive processes or the circulation of blood through the arteries. Is there a general characteristic that distinguishes mental phenomena from nonmental, or "merely" physical, phenomena? We canvass some suggestions for answering these questions later in this chapter.

There are also problems concerning specific mental properties or kinds of mental states and their relationship to one another. Are pains only sensory events (they hurt), or must they also have a motivational component (such as aversiveness)? Can there be pains of which we are not aware? Do emotions like anger and jealousy necessarily involve felt qualities? Do they involve a cognitive component, like belief? What is a belief anyway, and how does a belief come to have the content it has (say, that it is raining outside, or that 7 + 5 = 12)? Do beliefs and thoughts require a capacity for speech?

A third group of problems concerns the relation between minds and bodies, or between mental and physical phenomena. Collectively called "the mind-body problem," this has been a central problem of philosophy of mind since Descartes introduced it nearly four centuries ago. It is a central problem for us in this book as well. The task here is to clarify and make intelligible the relation between our mentality and the physical nature of our being—or more generally, the relationship between mental and physical properties. But why should we think there is a philosophical problem here? Just what needs to be clarified and explained?

A simple answer might go like this: The mental seems prima facie so utterly different from the physical, and yet the two seem intimately related to each other. When you think of conscious experiences—such as the smell of basil, a pang of remorse, or the burning painfulness of a freshly bruised elbow—it is hard to imagine anything that could be more different from mere configurations and motions, however complex, of material particles, atoms and molecules, or mere physical changes involving cells and tissues. In spite of that, these conscious phenomena don't come out of thin air, or from some immaterial source; rather, they arise from certain configurations of physical-biological processes of the body, including neural processes in the brain. We are at bottom physical-biological systems—complex biological structures wholly made up of bits of matter. (In case you disagree, we consider Descartes's contrary

views in chapter 2.) How can biological-physical systems come to have states like thoughts, fears, and hopes, experience feelings like guilt and pride, act for reasons, and be morally responsible? It strikes many of us that there is a fundamental, seemingly unbridgeable gulf between mental and physical phenomena and that this makes their apparently intimate relationships puzzling and mysterious.

It seems beyond doubt that phenomena of the two kinds are intimately connected. For one thing, evidence indicates that mental events occur as a result of physical-neural processes. Stepping barefoot on an upright thumbtack causes a sharp pain in your foot. It is likely that the proximate basis of the pain is some event in your brain: A bundle of neurons deep in your hypothalamus or cortex discharges, and as a result you experience a sensation of pain. Impingement of photons on your retina starts off a chain of events, and as a result you have a certain visual experience, which in turn leads you to form the belief that there is a tree in front of you. How could a series of physical events—little particles jostling against one another, electric current rushing to and fro, and so on—blossom all of a sudden into a conscious experience, like the burning hurtfulness of a badly scalded hand, the brilliant red and purple sunset you see over the dark green ocean, or the smell of freshly mown lawn? We are told that when certain special neurons (nociceptive neurons) fire, we experience pain, and presumably there is another group of neurons that fire when we experience an itch. Why are pain and itch not switched around? That is, why is it that we feel pain, rather than itch, when just these neurons fire and we experience itch, not pain, when those other neurons fire? Why is it not the other way around? Why should any experience emerge from molecular-biological processes?

Moreover, we take it for granted that mental events have physical effects. It seems essential to our concept of ourselves as agents that our bodies are moved in appropriate ways by our wants, beliefs, and intentions. You see a McDonald's sign across the street and you decide to get something to eat, and somehow your perception and decision cause your limbs to move in such a way that you now find your body at the doors of the restaurant. Cases like this are among the familiar facts of life and are too boring to mention. But how did your perception and desire manage to move your body, all of it, across the street? You say, that's easy: Beliefs and desires first cause certain neurons in the motor cortex of my brain to discharge, these neural impulses are transmitted through the network of neural fibers all the way down to the peripheral control systems, which cause the appropriate muscles to contract, and so on. All that might be a complicated story, you say, but it is something that

brain science, not philosophy, is in charge of explaining. But how do beliefs and desires manage to cause those little neurons to fire to begin with? How can this happen unless beliefs and desires are themselves just physical happenings in the brain? But is it coherent to suppose that these mental states are simply physical processes in the brain? These questions do not seem to be questions that can be answered just by doing more research in neuroscience; they seem to require philosophical reflection and analysis beyond what we can learn from science alone. This is what is called the problem of mental causation, one of the most important issues concerning the mind ever since Descartes first formulated the mind-body problem.

In this book, we are chiefly, though not exclusively, concerned with the mind-body problem. We begin, in the next chapter, with an examination of Descartes's mind-body dualism—a dualism of material things and immaterial minds. In contemporary philosophy of mind, however, the world is conceived to be fundamentally material: There are persuasive (some will say compelling) reasons to believe that the world we live in is made up wholly of material particles and their structured aggregates, all behaving strictly in accordance with physical laws. How can we accommodate minds and mentality in such an austerely material world? That is our main question.

But before we set out to consider specific doctrines concerning the mind-body relationship, it will be helpful to survey some of the basic concepts, principles, and assumptions that guide the discussions to follow.

Metaphysical Preliminaries

For Descartes, "having a mind" had a literal meaning. On his view, minds are things of a special kind, souls or immaterial substances, and having a mind simply amounts to having a soul, something outside physical space, whose essence consists in mental activities like thinking and being conscious. (We examine this view of minds in chapter 2.) A substantival view of mentality like Descartes's is not widely accepted today. However, to reject minds as substances or objects in their own right is not to deny that each of us "has a mind"; it is only that we need not think of "having a mind" as there being some object called a "mind" that we literally "have." Having a mind need not be like having brown eyes or a laptop. Think of "dancing a waltz" or "taking a walk": When we say, "Sally danced a waltz," or "Sally took a leisurely walk along the river," we do not mean—at least we do not need to mean—that there are *things* in this world called "waltzes" or "walks" such that Sally picked out one of them and danced it

or walked it. Where are these dances and walks when no one is dancing or walking them? What could you do with a dance except dance it? Dancing a waltz is not like owning an SUV or kicking a tire. Dancing a waltz is merely a *manner* of dancing, and taking a walk is a *manner* of moving your limbs in a certain relationship to the physical surroundings. In using these expressions, we need not accept the existence of entities like waltzes and walks; all we need to admit into our ontology—the scheme of entities we accept as real—are persons who waltz and persons who walk.

Similarly, when we use expressions like "having a mind," "losing one's mind," "being out of one's mind," and the like, there is no need to suppose there are objects in this world called "minds" that we have, lose, or are out of. Having a mind can be construed simply as having a certain group of *properties*, *features*, and *capacities* that are possessed by humans and some higher animals but absent in things like rocks and trees. To say that some creature "has a mind" is to classify it as a certain sort of being, capable of certain characteristic sorts of behaviors and functions—sensation, perception, memory, learning, reasoning, consciousness, action, and the like. It is less misleading, therefore, to speak of "mentality" than of "having a mind"; the surface grammar of the latter abets the problematic idea of a substantival mind—mind as an object of a special kind. However, this is not to preclude substantival minds at the outset; the point is only that we should not infer their existence from our use of certain forms of expression. As we will see in the chapter to follow, there are serious philosophical arguments that we must accept minds as immaterial things. Moreover, an influential contemporary view identifies minds with brains (discussed in chapter 4). Like Descartes's substance dualism, this view gives a literal meaning to "having a mind": It would simply mean having a brain of certain structure and capacities. The main point we should keep in mind is that all this requires philosophical considerations and arguments, as we will see in the rest of this book.

Mentality is a broad and complex category. As we just saw, there are numerous specific properties and functions through which mentality manifests itself, such as experiencing sensations, entertaining thoughts, reasoning and judging, making decisions, and feeling emotions. There are also more specific properties that fall within these categories, such as experiencing a throbbing pain in the right elbow, believing that Kabul is in Afghanistan, wanting to visit Tibet, and being annoyed at your roommate. In this book, we often talk in terms of "instantiating," "exemplifying," or "having" this or that property. When you shut a door on your thumb, you will likely *instantiate* or *exemplify*

the property of being in pain; most of us *have*, or *instantiate*, the property of believing that snow is white; some of us have the property of wanting to visit Tibet; and so on. Admittedly this is a somewhat cumbersome, not to say stilted, way of talking, but it gives us a uniform and simple way of referring to certain entities and their relationships. Throughout this book, the expressions "mental" and "psychological" and their respective cognates are used interchangeably. In most contexts, the same goes for "physical" and "material."

We will now set out in general terms the kind of ontological scheme that we presuppose in this book and explain how we use certain terms associated with the scheme. We suppose, first, that our scheme includes *substances*, that is, *things* or *objects* (including persons, biological organisms and their organs, molecules, computers, and such) and that they have various *properties* and stand in various *relations* to each other. (Properties and relations are together called *attributes*.) Some of these are physical, like having a certain mass or temperature, being one meter long, being longer than, and being between two other objects. Some things—in particular, persons and certain biological organisms—can also instantiate mental properties, like being in pain, fearing darkness, and disliking the smell of ammonia. We also speak of mental or physical *events*, *states*, and *processes* and sometimes of *facts*. A process can be thought of as a (causally) connected series of events and states; events differ from states in that they suggest *change*, whereas states do not. The terms "phenomenon" and "occurrence" can be used to cover both events and states. We often use one or another of these terms in a broad sense inclusive of the rest. (For example, when we say "every event has a cause," we are not excluding states, phenomena, and the rest.) How events and states are related to objects and their properties is a question of some controversy in metaphysics. We simply assume here that when a person instantiates, at time *t*, a mental property—say, being in pain—then there is the event (or state) of that person's being in pain at *t*, and there is also the fact that the person is in pain at *t*. Some events are psychological events, such as pains, beliefs, and onsets of anger, and these are instantiations by persons and other organisms of mental properties. Some events are physical, such as earthquakes, hiccups and sneezes, and the firing of a bundle of neurons, and these are instantiations of physical properties. Another point to note: In the context of the mind-body problem, the physical usually goes beyond the properties and phenomena studied in physics; the biological, the chemical, the geological, and so on, also count as physical.

So much for the ontological preliminaries. Sometimes clarity and precision demand attention to ontological details, but as far as possible we will try to

avoid general metaphysical issues that are not germane to our concerns about the nature of mind.

Mind-Body Supervenience

Consider the apparatus called the "transporter" in the science-fiction television series *Star Trek*. You walk into a booth. When the transporter is activated, your body is instantly disassembled; exhaustive information concerning your bodily structure and composition, down to the last molecule, is transmitted, instantaneously, to another location, often a great distance away, where a body that is exactly like yours is reconstituted (presumably with local material). And someone who looks just like you materializes on the spot and starts doing the tasks you were assigned to do there.

Let us not worry about whether the person who is created at the destination is really you or only your replacement. In fact, we can avoid this issue by slightly changing the story: Exhaustive information about your bodily composition is obtained by a scanner that does no harm to you, and on the basis of this information, an exact physical replica of your body—a molecule-for-molecule identical duplicate—is created at another location. By assumption, you and your replica have exactly the same *physical* properties; you and your replica could not be distinguished by any *current intrinsic* physical differences. We say "current" to rule out the obvious possibility of distinguishing you from your duplicate by tracing the causal chains backward to the past. We say "intrinsic" because you and your replica have different relational, or extrinsic, properties; for example, you have a mother but your replica does not.

Given that your replica is your *physical* replica, will she also be your *psychological* replica? Will she be identical with you in all mental respects as well? Will she be as smart and witty as you are, and as prone to daydream? Will she share your likes and dislikes in food and music and behave just as you would when angry or irritable? Will she prefer blue to green and have a visual experience exactly like yours when you and she both gaze at a Van Gogh landscape of yellow wheat fields against a dark blue sky? Will her twinges, itches, and tickles feel to her just the way yours feel to you? Well, you get the idea. An unquestioned assumption of *Star Trek* and similar science-fiction fantasies seems to be that the answer is yes to each of these questions. If you are like the many *Star Trek* fans in going along with this assumption, that is because you have tacitly consented to the following "supervenience" thesis:

> *Mind-Body Supervenience I.* The mental supervenes on the physical in that things (objects, events, organisms, persons, and so on) that are exactly alike in all physical properties cannot differ with respect to mental properties. That is, physical indiscernibility entails psychological indiscernibility.

Or as it is sometimes put: No mental difference without a physical difference. Notice that this principle does not say that things that are alike in psychological respects must be alike in physical respects. We seem to be able coherently to imagine intelligent extraterrestrial creatures whose biochemistry is different from ours (say, their physiology is not carbon-based) and yet who share the same psychology with us. As we might say, the same psychology could be realized in different physical systems. Now, that may or may not be the case. The thing to keep in mind, though, is that mind-body supervenience asserts only that creatures could not be psychologically different and yet physically identical.

There are two other important ways of explaining the idea that the mental supervenes on the physical. One is the following, known as "strong supervenience":

> *Mind-Body Supervenience II.* The mental supervenes on the physical in that if anything *x* has a mental property M, there is a physical property P such that *x* has P, and necessarily any object that has P has M.

Suppose that a creature is in pain (that is, it has the mental property of being in pain). This supervenience principle tells us that in that case there is some physical property P that the creature has that "necessitates" its being in pain. That is to say, pain has a physical substrate (or "supervenience base") such that anything that has this underlying physical property must be in pain. Thus, this formulation of mind-body supervenience captures the idea that the instantiation of a mental property in something "depends" on its instantiating an appropriate physical "base" property (that is, a neural correlate or substrate). How is this new statement of mind-body supervenience related to the earlier statement? It is pretty straightforward to show that the supervenience principle (II) entails (I); that is, if the mental supervenes on the physical according to (II), it will also supervene according to (I). Whether (I) entails (II) is more problematic.[1] For practical purposes, however, the two principles may be considered equivalent, and we make use of them in this book without worrying about their subtle differences.

There is another common way of understanding the supervenience relationship:

> *Mind-Body Supervenience III.* The mental supervenes on the physical in that worlds that are alike in all physical respects are alike in all mental respects as well; in fact, worlds that are physically alike are exactly alike overall.[2]

This formulation of supervenience, called "global" supervenience, states that if there were another world that is just like our world in all physical respects, with the same particles, atoms, and molecules in the same places and the same laws governing their behavior, the two worlds could not differ in any mental respects. If God created this world, all he had to do was to put the right basic particles in the right places and fix basic physical laws, and all else, including all aspects of mentality, would just come along. Once the basic physical structure is put in place, his job is finished; he does not *also* have to create minds or mentality, any more than trees or mountains or bridges. The question whether this formulation of supervenience is equivalent to either of the earlier two is a somewhat complicated one; let it suffice to say that there are close relationships between all three. In this book, we do not have an occasion to use (III); however, it is stated here because this is the formulation some philosophers favor and you will likely come across it in the philosophy of mind literature.

To put mind-body supervenience in perspective, it might be helpful to look at supervenience theses in other areas—in ethics and aesthetics. Most moral philosophers would accept the thesis that the ethical, or normative, properties of persons, acts, and the like are supervenient on their nonmoral, descriptive properties. That is, if two persons, or two acts, are exactly alike in all nonmoral respects (say, the persons are both honest, courageous, kind, generous, and so on), they could not differ in moral respects (say, one of them is a morally good person but the other is not). Supervenience seems to apply to aesthetic qualities as well: If two pieces of sculpture are physically exactly alike (the same shape, size, color, texture, and all the rest), they cannot differ in some aesthetic respect (say, one of them is elegant, heroic, and expressive while the second has none of these properties). A world molecule-for-molecule identical with our world will contain works of art just as beautiful, noble, and mysterious as our Michelangelos, Vermeers, and Magrittes. One more example: Just as mental properties are thought to supervene on physical properties, most consider biological properties to supervene on more basic physicochemical properties. It seems natural to suppose that if two things are exactly alike in basic physical

and chemical features, including, of course, their material composition and structure, it could not be the case that one of them is a living thing and the other is not, or that one of them is performing a certain biological function (say, photosynthesis) and the other is not. That is to say, physicochemically indiscernible things must be biologically indiscernible.

As noted, most philosophers accept these supervenience theses; however, whether they are true, or why they are true, are philosophically nontrivial questions. And each supervenience thesis must be evaluated and assessed on its own merit. Mind-body supervenience, of course, is our present concern. Our ready acceptance of the idea of the *Star Trek* transporter shows the strong intuitive attraction of mind-body supervenience. But is it true? What is the evidence in its favor? Should we accept it? These are deep and complex questions. One reason is that, in spirit and substance, they amount to the following questions: Is physicalism true? Should we accept physicalism?

Materialism and Physicalism

Since materialism, or physicalism, broadly understood is the basic framework in which contemporary philosophy of mind has been debated, it is useful for us to begin with some idea of what it is. Materialism is the doctrine that all things that exist in the world are bits of matter or aggregates of bits of matter. There is no thing that isn't a material thing—no transcendental beings, Hegelian "absolutes," or immaterial minds. Physicalism is the contemporary successor to materialism. The thought is that the traditional notion of material stuff was ill-suited to what we now know about the material world from contemporary physics. For example, the concept of a "field" is widely used in physics, but it is unclear whether fields would count as material things in the traditional sense. Physicalism is the doctrine that all things that exist are entities recognized by the science of physics, or systems aggregated out of such entities.[3] According to some physicalists, so-called nonreductive physicalists, these physical systems can have nonphysical properties, properties that are not recognized by physics or reducible to them. Psychological properties are among the prime candidates for such nonphysical properties possessed by physical systems.

If you are comfortable with the idea of the *Star Trek* transporter, that means you are comfortable with physicalism as a perspective on the mind-body problem. The wide and seemingly natural acceptance of the transporter idea shows how pervasively physicalism has penetrated contemporary culture, although when this is made explicit some people would no doubt recoil and proclaim themselves to be against physicalism.

What is the relationship between mind-body supervenience and physicalism? We have not so far defined what physicalism is, but the term itself suggests that it is a doctrine that affirms the primacy, or basicness, of what is physical. With this very rough idea in mind, let us see what mind-body supervenience implies for the dualist view (to be discussed in more detail in chapter 2) associated with Descartes that minds are immaterial substances with no physical properties whatever. Take two immaterial minds: Evidently, they are exactly alike in all physical respects since neither has any physical property and as a result it is impossible to distinguish them from a physical perspective. So if mind-body supervenience, in the form of (I), holds, it follows that they are alike in all mental respects. That is, under mind-body supervenience (I), all Cartesian immaterial souls are exactly alike in all mental respects, from which it follows that they are exactly alike in all possible respects. From this it seems to follow that there can be at most one immaterial soul! No serious mind-body dualist would find these consequences of mind-body supervenience tolerable. This is one way of seeing why the dualist will want to reject mind-body supervenience.

To appreciate the physicalist implication of mind-body supervenience, we must consider one aspect of supervenience that we have not so far discussed. Many philosophers regard the supervenience thesis as affirming a relation of *dependence* or *determination* between the mental and the physical; that is, the mental properties a given thing has depend on, or are determined by, the physical properties it has. Consider version (II) of mind-body supervenience: It says that for every mental property M, if anything has M, it has some physical property P that *necessitates* M—if anything has P, it *must* have M. This captures the idea that mental properties must have neural, or other physical, "substrates" from which they arise and that there can be no instantiation of a mental property that is not grounded in some physical property. So a dependence relation can naturally be read into the claim that the mental supervenes on the physical, although, strictly speaking, the supervenience theses as stated only make claims about how mental properties *covary* with physical properties. In any case, many physicalists interpret supervenience as implying mind-body dependence in something like the following sense:

> *Mind-Body Dependence*. The mental properties a given thing has depend on, and are determined by, the physical properties it has. That is, our psychological character is wholly determined by our physical nature.

The dependence thesis is important because it is an explicit affirmation of the *ontological primacy*, or *priority*, of the physical in relation to the mental. The

thesis seems to accord well with the way we ordinarily think of the mind-body relation, as well as with scientific assumptions and practices. Few of us would think that there can be mental events and processes that float free, so to speak, of physical processes; most of us believe that what happens in our mental life, including the fact that we have a mental life at all, is dependent on what happens in our body, in particular in our nervous system. Furthermore, it is because mental states depend on what goes on in the brain that it is possible to intervene in the mental goings-on. To ease your headache, you take aspirin—the only way you can affect the headache is to alter the neural base on which it supervenes. There apparently is no other way.

For these reasons, we can think of the mind-body supervenience thesis, in one form or another, as *minimal physicalism*, in the sense that it is one commitment that all who consider themselves physicalists must accept. But is it sufficient as physicalism? That is, can we say that anyone who accepts mind-body supervenience is ipso facto a full physicalist? Opinions differ on this question. We saw earlier that supervenience does not by itself completely rule out the existence of immaterial minds, something antithetical to physicalism. But we also saw that supervenience has consequences that no serious dualist can accept. Whether supervenience itself suffices to deliver physicalism depends, by and large, on what we consider to be full and robust physicalism. As our starting options, then, let us see what varieties of physicalism are out there.

First, there is an ontological claim about what objects there are in this world:

> *Substance Physicalism*.[4] All that exists in this world are bits of matter in space-time and aggregate structures composed of bits of matter. There is nothing else in the space-time world.

This thesis, though it is disputed by Descartes and other substance dualists, is accepted by most contemporary philosophers of mind. The main point of contention concerns the *properties* of material or physical things. Certain complex physical systems, like higher organisms, are also psychological systems; they exhibit psychological properties and engage in psychological activities and functions. How are the psychological properties and physical properties of a system related to each other? Broadly speaking, an ontological physicalist has a choice between the following two options:

> *Property Dualism, or Nonreductive Physicalism*. The psychological properties of a system are distinct from, and irreducible to, its physical properties.[5]

> *Reductive Physicalism*, or *Type Physicalism*. Psychological properties (or kinds, types) are reducible to physical properties (kinds, types). That is, psychological properties and kinds are physical properties and kinds. There are only properties of one sort exemplified in this world, and they are physical properties.

Remember that for our purposes "physical" properties include chemical, biological, and neural properties, not just those properties investigated in basic physics (such as energy, mass, or charm). You could be a property dualist because you reject mind-body supervenience, but then you would not count as a physicalist since, as we argued, mind-body supervenience is a necessary element of physicalism. So the physicalist we have in mind is someone who accepts mind-body supervenience. However, it is generally supposed that mind-body supervenience is consistent with property dualism, the claim that the supervenient psychological properties are irreducible to, and not identical with, the underlying physical base properties. In defense of this claim, some point to the fact that philosophers who accept the supervenience of moral properties on nonmoral, descriptive properties for the most part reject the reducibility of moral properties, like being good or being right, to nonmoral, purely descriptive properties. The situation seems the same with the case of aesthetic supervenience and aesthetic properties.[6]

Some philosophers who reject reductive, or type, physicalism as too ambitious and overreaching embrace "token" physicalism—the thesis that although psychological types are not identical with physical types, each and every individual psychological event, or event-token, is a physical event. So pain, as a mental kind, is not identical with, or reducible to, a kind of physical event or state, and yet each individual instance of pain—this pain here now—is usually a physical event. Token physicalism is considered a form of nonreductive physicalism. The continuing debate between nonreductive physicalists and reductive physicalists has largely shaped the contemporary debate on the mind-body problem.[7]

Varieties of Mental Phenomena

It is useful at this point to look at some major categories of mental events and states. This will give us a rough idea about the kinds of phenomena we are concerned with and also remind us that the phenomena that come under the rubric "mental" or "psychological" are extremely diverse and variegated. The following list is not intended to be complete or systematic, and some categories obviously overlap others.

First, we may distinguish those mental phenomena that involve *sensations* or *sensory qualities*: pains, itches, tickles, having an afterimage, seeing a round green patch, smelling ammonia, feeling nauseous, and so on. These mental states are said to have a "phenomenal" or "qualitative" character—the way they *feel* or the way they *look* or *appear*. To use a popular term, there is *something it is like* to experience such phenomena or be in such states. Thus, pains have a special qualitative feel that is distinctive of pains—they hurt. Similarly, itches itch and tickles tickle. When you look at a green patch, there is a distinctive way the patch looks to you: It looks *green*, and your visual experience involves this green look. Each such sensation has its own distinctive feel and is characterized by a sensory quality that we seem to be able to identify directly, at least as to the general type to which it belongs (for example, pain, itch, or seeing green). These items are called "phenomenal" or "qualitative states," or sometimes "raw feels." However, "qualia" has now become the standard term for these sensory, qualitative states, or the sensory qualities experienced in such states. Collectively, these mental phenomena are said to constitute "phenomenal consciousness."

Second, there are mental states that are attributed to a person by the use of embedded that-clauses: for example, President Barack Obama *hopes* that Congress will pass a health-care bill this year; Senator Harry Reid *is certain* that this will happen, and Newt Gingrich doubts that Obama will get what he wants. Such states are called "propositional attitudes." The idea is that these states consist in a subject's having an "attitude" (for example, hoping, being certain, doubting, and believing) toward a "proposition" (for example, that Congress will pass a health-care bill, that it will rain tomorrow). The propositions are said to constitute the "content" of the propositional attitudes, and that-clauses that specify these propositions are called "content sentences." Thus, the content of Obama's hope is the proposition that Congress will pass a health-care bill this year, which is also the content of Gingrich's doubt, and this content is expressed by the sentence "Congress will pass a health-care bill this year." These states are also called "intentional"[8] or "content-bearing" states.

Do these mental states have a phenomenal, qualitative aspect? We do not normally associate a specific feel with beliefs, another specific feel with desires, and so on. There does not seem to be any special belief-like feel, a common sensory quality, associated with your belief that Providence is south of Boston and your belief that two is the smallest prime number. At least it seems that we can say this much: If you believe that two is the smallest prime and I do too, there does not seem to be—nor need there be—any common sensory quality that both of us experience in virtue of sharing this belief. The importance of these

intentional states cannot be overstated. Much of our ordinary psychological thinking and theorizing ("commonsense" or "folk" psychology) involves propositional attitudes; we make use of them all the time to explain and predict what people will do. Why did Mary cross the street? Because she wanted some coffee and thought that she could get it at the Starbucks across the street. These states are essential to social psychology, and their analogues are found in various areas of psychology and cognitive science.

And then there are various mental states that come under the broad heading of *feelings* and *emotions*. They include anger, joy, sadness, depression, elation, pride, embarrassment, remorse, regret, shame, and many others. Notice that emotions are often attributed to persons with a that-clause. In other words, some states of emotions involve propositional attitudes: For example, you could be *embarrassed* that you had forgotten to call your mother on her birthday, and she could be *disappointed* that you did. Further, some emotions involve belief: If you are embarrassed that you had forgotten your mother's birthday, you must believe that you did. As the word *feeling* suggests, there is often a special qualitative component we associate with many emotions, such as anger and grief, although it is not certain that all instances of emotion are accompanied by such qualitative feels, or that there is a single specific sensory feel to each kind of emotion.

There are also what some philosophers call "volitions," like intending, deciding, and willing. These states are propositional attitudes; intentions and decisions have content. For example, I may intend to take the ten o'clock train to New York tomorrow; here the content is expressed by an infinitive construction ("to take"), but it is easily spelled out in a full sentence, as in "I intend that I take the ten o'clock train to New York tomorrow." In any case, these states are closely related to actions. When I intend to raise my arm *now*, I must *now* undertake to raise my arm; when you intend, or decide, to do something, you commit yourself to doing it. You must be prepared not only to take the necessary steps toward doing it but also to initiate them at an appropriate time. This is not to say that you cannot change your mind, or that you will necessarily succeed; it is to say that you need to change your intention to be released from the commitment to action. According to some philosophers, all intentional actions must be preceded by an act of volition.

Actions typically involve motions of our bodies, but they do not seem to be mere bodily motions. My arm is going up, and so is yours. However, you are raising your arm, but I am not—my arm is being pulled up by someone else. The raising of your arm is an action; it is something you do. But the rising of my arm is not an action; it is not something that I do but something that happens

to me. There appears to be something mental about your raising your arm that is absent from the mere rising of an arm; perhaps it is the involvement of your desire, or intention, to raise your arm, but exactly what distinguishes actions from "mere bodily motions" has been a matter of philosophical dispute. Or consider something like buying a loaf of bread. Evidently someone who can engage in the act of buying a loaf of bread must have appropriate beliefs and desires; she must, for example, have a desire to buy bread, or at least a desire to buy something, and knowledge of what bread is. And to do something like buying, you must have knowledge, or beliefs, about what constitutes buying rather than, say, borrowing or simply taking, about money and exchange of goods, and so on. That is to say, only creatures with beliefs and desires and an understanding of appropriate social conventions and institutions can engage in activities like buying and selling. The same goes for much of what we do as social beings; actions like promising, greeting, and apologizing presuppose a rich and complex background of beliefs, desires, and intentions, as well as an understanding of social relationships and arrangements.

There are other items that are ordinarily included under the rubric of "psychological," such as traits of character and personality (being honest, obsessive, witty, introverted), habits and propensities (being industrious, punctual), intellectual abilities, artistic talents, and the like. But we can consider them to be mental in an indirect or derivative sense: Honesty is a mental characteristic because it is a tendency, or disposition, to form desires of certain sorts (for example, the desire to tell the truth, or not to mislead others) and to act in appropriate ways (in particular, saying only what you sincerely believe).

In the chapters to follow, we focus on sensations and intentional states. They provide us with examples of mental states when we discuss the mind-body problem and other issues. We also discuss some specific philosophical problems about these two principal types of mental states. We will largely bypass detailed questions, however, such as what types of mental states there are, how they are interrelated, and the like.

But in what sense are all these variegated items "mental" or "psychological"? Is there some single property or feature, or a reasonably simple and perspicuous set of them, by virtue of which they all count as mental?

Is There a "Mark of the Mental"?

Various characteristics have been proposed by philosophers to serve as a "mark of the mental," a criterion that would separate mental phenomena or properties from those that are not mental. Each has a certain degree of plausibility and can

be seen to cover a range of mental phenomena, but as we will see, none seems to be adequate for all the diverse kinds of events and states, functions and capacities, that we normally classify as "mental" or "psychological." Although we will not try to formulate our own criterion of the mental, a review of some of the prominent proposals will give us an understanding of the principal ideas traditionally associated with the concept of mentality and highlight some of the important characteristics of mental phenomena, even if, as noted, no single one of them seems capable of serving as a universal, necessary, and sufficient condition of mentality.

Epistemological Criteria

You are experiencing a sharp toothache caused by an exposed nerve in a molar. The toothache that you experience, but not the condition of your molar, is a mental occurrence. But what is the basis of this distinction? One influential answer says that the distinction consists in certain fundamental differences in the way you come to have *knowledge* of the two phenomena.

Direct or Immediate Knowledge. Your knowledge that you have a toothache, or that you are hoping for rain tomorrow, is "direct" or "immediate"; it is not based on evidence or inference. There is nothing else that you know or need to know from which you infer that you have a toothache; that is, your knowledge is not mediated by other beliefs or knowledge. This is seen in the fact that in cases like this the question "How do you know?" seems to be out of place ("How do you know you are hoping for rain and not snow?"). The only possible answer, if you take the question seriously, is that you *just* know. This shows that here the question of "evidence" is inappropriate: Your knowledge is direct and immediate, not based on evidence. Yet your knowledge of the physical condition of your tooth is based on evidence: Knowledge of this kind usually depends on the testimonial evidence provided by a third party—for example, your dentist. And your dentist's knowledge presumably depends on the evidence of X-rays, visual inspection of your teeth, and so on. The question "How do you know that you have an exposed nerve in a molar?" makes good sense and can receive an informative answer.

But isn't our knowledge of certain simple physical facts just as "direct" and "immediate" as knowledge of mental events like toothaches and itches? Suppose you are looking at a large red circle painted on a wall directly in front of you: Doesn't it seem that you know, directly and without the use of any further evidence, that there is a round red patch in front of you? Don't I know, in

the same way, that here is a piece of white paper in front of me or that there is a tree just outside my window?

Privacy, or First-Person Privilege. One possible response to the foregoing challenge is to invoke the privacy of our knowledge of our own mental states, namely, the apparent fact that this direct access to a mental event is enjoyed by a single subject, the person to whom the event is occurring. In the case of the toothache, it is only you, not your dentist or anyone else, who is in this kind of specially privileged position. But this does not hold in the case of seeing the red patch. If you can know "directly" that there is a round red spot on the wall, so can I and anyone else who is suitably situated in relation to the wall. There is no single person with specially privileged access to the round red spot. In this sense, knowledge of mental events exhibits an *asymmetry* between first person and third person: It is only the first person, namely the subject who experiences a pain, who enjoys a special epistemic privilege as regards the pain. Others, that is, third persons, do not. In contrast, for knowledge of physical objects and states—say, the red round spot on the wall—there is no meaningful first-person/third-person distinction; everyone is a third person. Moreover, the first-person privilege holds only for knowledge of *current* mental occurrences, not for knowledge of *past* ones: You know that you had a toothache yesterday, a week ago, or two years ago, from the evidence of memory, an entry in your diary, your dental record, and the like.

But what about those bodily states we detect through proprioception, such as the positions and motions of our limbs (for example, knowing that your legs are crossed or that you are raising your right hand)? Our proprioceptors and associated neural machinery are in the business of keeping us *directly* informed of certain physical conditions of our bodies, and proprioception is, in general, highly reliable. Moreover, first-person privilege seems to hold for such cases: It is only I who know, through proprioception, that my right knee is bent; no third party has similar access to this fact. And yet it is knowledge of a bodily condition, not of a mental occurrence. Perhaps this example could be handled by appealing to the following criterion.

Infallibility and Transparency. Another epistemic feature sometimes associated with mentality is the idea that in some sense your knowledge of your own current mental states is "infallible" or "incorrigible," or that it is "self-intimating" (or that your mind is "transparent" to you). The main idea is that mental events—especially events like pains and other sensations—have the following property: You cannot be mistaken about whether you are experiencing

them. That is, if you *believe* that you are in pain, then it follows that you *are* in pain, and if you believe that you are not in pain, then you are not; it is not possible to have false beliefs about your own pains. In this sense, your knowledge of your own pain is *infallible*. So-called psychosomatic pains are pains nonetheless; they can hurt just as badly. The same may hold for your knowledge of your own propositional attitudes like belief; Descartes famously said that you cannot be mistaken about the fact that you doubt, or that you think.[9] In contrast, when your belief concerns a physical occurrence, there is no guarantee that your belief is true: Your belief that you have a decayed molar may be true, but its truth is not entailed by the mere fact that you believe it. Or so goes the claim, at any rate. Returning briefly to knowledge gained through proprioception, the reply would be that such knowledge may be reliable but not infallible; there can be incorrect beliefs about your bodily position based on proprioception.

Transparency is the converse of infallibility: A state or event *m* is said to be transparent to a person just in case, necessarily, if *m* occurs, the person is aware that *m* occurs—that is, she knows that *m* occurs.[10] The claim, then, is that mental events are transparent to the subjects to whom they occur. If pains are transparent in this sense, there could not be *hidden* pains—pains that the subject is unaware of. Just as the infallibility of beliefs about your own pains implies that pains with no physiological cause are at least conceivable, the transparent character of pain implies that even if all normal physical and physiological causes of pain are present, if you are not aware of any pain, then you are not in pain. There are reports about soldiers in combat and athletes in the heat of competition that they experienced no pain in spite of severe physical injuries; if we assume pains are transparent, we would have to conclude that pain, as a mental event, is not occurring to these subjects. We may define "the doctrine of the transparency of mind" as the claim that nothing that happens in your mind escapes your awareness—that is, nothing in your mind is hidden from you. The conjunction of this doctrine and the doctrine of infallibility is often associated with the traditional conception of the mind, especially that of Descartes.

Infallibility and transparency are extremely strong properties. It would be no surprise if physical events and states did not have them; a more interesting question is whether all or even most mental events satisfy them. Evidently, not all mental events or states have these special epistemic properties. In the first place, it is now commonplace to speak of "unconscious" or "subconscious" beliefs, desires, and emotions, like repressed desires and angers—psychological states the subject is not aware of and would even vehemently deny having but that evidently shape and influence his action and behavior. Second, it is not al-

ways easy for us to determine whether an emotion that we are experiencing is, say, one of embarrassment, remorse, or regret—or one of envy, jealousy, or anger. And we are often not sure whether we "really" believe or desire something. Do I believe that globalization is a good thing? Do I believe that I am by and large a nice person? Do I want to be sociable and gregarious, or do I prefer to stay somewhat aloof and distant? If you reflect on such questions, you may not be sure what the answers are. It is not as though you have suspended judgment about them—you may not even know *that*. Epistemic uncertainties can happen with sensations as well. Does this overripe avocado smell a little like a rotten egg, or is it okay for the salad? Special epistemic access is perhaps most plausible for sensations like pains and itches, but here again, not all our beliefs about pains appear to have the special authoritative character indicated by the epistemic properties we have surveyed. Is the pain I am now experiencing more intense than the pain I felt a moment ago in the same elbow? Just where in my elbow is the pain? Clearly there are many characteristics of pains, even introspectively identifiable ones, about which I could be mistaken and don't feel fully certain.

It is also thought that you can misclassify, or misidentify, the type of sensation you are experiencing: For example, you may report that you are itchy in the shoulder when the correct description would be that you have a ticklish sensation there. However, it is not clear just what such cases show. It might be replied, for example, that the error is a verbal one, not one of belief. Although you are using the sentence "My left shoulder is itchy" to report your belief, your belief is to the effect that your left shoulder *tickles*, and this belief is true.

Thus, exactly how the special epistemic character of mental events is to be characterized can be a complex and controversial business, and unsurprisingly there is little agreement in this area. Some philosophers, especially those who favor a scientific approach to mentality, would take pains to minimize these prima facie differences between mental and bodily events. But it is apparent that there are important epistemological differences between the mental and the nonmental, however the differences are to be precisely described. Especially important is the first-person epistemic authority noted earlier: We seem to have special access to our own mental states—or at least to an important subclass of them if not all of them. Such access may well fall short of infallibility or incorrigibility, and it seems beyond doubt that our minds are not wholly transparent to us. But the differences we have noted, even if they are not quite the way described, are real enough, and they may be capable of serving as a starting point for thinking about our concepts of the mental and the physical. It may be that we get our initial purchase on the concept of mentality

through the core class of mental states for which some form of special first-person authority holds and that we derive the broader class of mental phenomena by extending and generalizing this core in various ways.[11]

Mentality as Nonspatial

For Descartes, the essential nature of a mind is that it is a thinking thing ("res cogitans"), and the essential nature of a material thing is that it is a spatially extended thing—something with a three-dimensional bulk. A corollary of this, for Descartes, is that the mental is essentially nonspatial and the material is essentially lacking in the capacity for thinking. Most physicalists would reject this corollary even if they accept the thesis that the mental is definable as thinking; they will say that as it happens, some material things, like higher biological organisms, can think, feel, and be conscious. But there may be a way of developing the idea that the mental is nonspatial that leaves the question of physicalism open.

For example, we might try something like this: To say that M is a mental property is to say that the proposition that something has M *does not logically imply* that it is a spatially extended thing. This allows the possibility that something that has M is in fact a spatially extended thing, though it is not required to be. So it may be that *as a matter of contingent fact*, all things that have mental properties are spatially extended things, like human beings and other biological organisms.

Thus, from the proposition that something *x* believes that four is an even number, it does not seem to follow that *x* is a spatially extended thing. There may be no immaterial angels in this world, but it does not seem logically contradictory to say that there are angels or that angels have beliefs and other mental states, like desires and hopes. But it evidently is a contradiction to say that something has a physical property—say, the color red, a triangular shape, or a rough texture—and at the same time to deny that it is something with spatial extensions. What about being located at a geometric point? Or *being* a geometric point, for that matter? But no physical *thing* is a geometric point; geometric points are not physical objects, and no physical object has the property of being a point or being located wholly at a point in space.

How useful is this nonspatiality approach toward a mark of the mental? It would seem that if you take this approach seriously, you must also take the idea of immaterial mental substance seriously. For you must allow the existence of possible worlds in which mental properties are instantiated by nonphysical beings (beings without spatial extension). The reasoning leading to this conclu-

sion is straightforward: Any mental property M is such that something can instantiate M without being spatially extended—that is, without being a physical thing. So M can be instantiated even if there is no physical thing. It follows then that there is a possible world in which mental properties, like belief and pain, are instantiated, even though no physical things exist in that world. What objects are there in such a world to serve as instantiators, or bearers, of mental properties? Since it makes no sense to think of abstract objects, like numbers, as possessors of mental properties, the only remaining possibility seems to be immaterial mental substances. It follows then that anyone who accepts the criterion of the mental as nonspatial must accept the idea of immaterial substance as a coherent one and allow the possible existence of such substances. This means that if you have qualms about the coherence of the Cartesian conception of minds as mental substances (see chapter 2), you would be well advised to stay away from the nonspatiality criterion of mentality.

Intentionality as a Criterion of the Mental

Schliemann sought the site of Troy. He was lucky; he found it. Ponce de León sought the Fountain of Youth, but he never found it. He could not have found it, since it does not exist and never did. It remains true, though, that he looked for the Fountain of Youth with great tenacity. The nonexistence of Bigfoot or the Loch Ness Monster has not prevented people from looking for them. Not only can you *look for* something that does not exist, but you can apparently also *think about*, *have beliefs and desires about*, *write about*, and even *worship* a nonexistent object. Even if God should not exist, he could be, and has been, the object of these mental acts or attitudes on the part of many people. Contrast these mental acts and states with physical ones, like cutting, kicking, and being to the left of. You cannot cut a nonexistent piece of wood, kick nonexistent tires, or be to the left of a nonexistent tree. That you kick something logically entails that the thing exists. That you are thinking of some object does not entail its existence. Or so it seems.

The Austrian philosopher Franz Brentano called this feature "the intentional inexistence" of psychological phenomena, claiming that it is this characteristic that separates the mental from the physical. In a famous passage, he wrote:

> Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as

> meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on.[12]

This feature of the mental—namely, that mental states are *about*, or are *directed upon*, objects that may or may not exist or have contents that may or may not be true—has been called "intentionality."

Broadly speaking, intentionality refers to the mind-world relation—specifically, the fact that our thoughts relate to, or hook up with, the things in the world, and represent how things are in the world. The idea at bottom is the thought that mentality is the capacity for representing the world around us, and that this is one of its essential functions. In short, the mind is a repository of inner representations—an inner mirror—of the outer world. The concept of intentionality may be subdivided into *referential intentionality* and *content intentionality*. Referential intentionality concerns the *aboutness* or *reference* of our thoughts, beliefs, intentions, and the like. When Ludwig Wittgenstein asked, "What makes my image of him into an image of *him*?"[13] he was asking for an explanation of what makes it the case that a given mental state (my "image of him") is *about*, or *refers to*, a particular object—him—rather than someone else. (That person may have an identical twin, and your image may fit his twin just as well, perhaps even better, but your image is of him, not of his twin. You may not even know that he has a twin.) Our words, too, refer to, or are directed upon, objects; "Mount Everest" refers to Mount Everest, and "horse" refers to horses.

Content intentionality concerns the fact that, as we saw, an important class of mental states—that is, propositional attitudes such as beliefs, hopes, and intentions—have contents or meanings, which are often expressed by full sentences. It is in virtue of having contents that our mental states *represent* states of affairs in the world. My perceiving that there are sunflowers in the field represents the fact, or state of affairs, of there being sunflowers in the field, and your remembering that there was a thunderstorm last night represents the state of affairs of there having been a thunderstorm last night. The capacity of our mental states to represent things external to them—that is, the fact that they have *representational content*—is clearly a very important fact about them. Obviously, our capacity to have representations of the outside world is critical to our ability to cope with our environment and survive and prosper. In

short, it is what makes it possible for us to have knowledge of the world. On a standard account, having knowledge is a matter of having mental representations with true contents—that is, representations that correctly represent.

Thus, referential intentionality and content intentionality are two related aspects of the fact that mental states have the capacity, and function, of representing things and states of affairs in the world. Brentano's thought seems to be that this representational capacity is the essence of the mind. It is the mind's essential function and raison d'être.

But can intentionality serve as the defining characteristic for all of mentality? Concerning the idea of representation, there is one point we must keep in mind: A representation has "satisfaction" conditions. In the case of representations with content intentionality, like the belief that snow is white, they can be evaluated in terms of truth or correctness. Pictorial or visual representations can be evaluated in terms of degrees of accuracy and fidelity. That means that a representation may fail to correctly represent—that is, it can misrepresent. In the case of referential intentionality, like "London" and "the Fountain of Youth," we can talk about their successfully referring to the intended object—an existing object. "London" refers to the city London, whereas "the Fountain of Youth" turned out to refer to nothing.

With this preliminary out of the way, there are two issues about intentionality as a criterion of the mental we need to discuss. The first is that some mental phenomena—in particular, bodily sensations like pains and tickles and orgasms[14]—do not seem to exhibit either kind of intentionality. The sensation of pain does not seem to be "about," or to refer to, anything; nor does it have a content that can be true or false, accurate or inaccurate. Doesn't the pain in my knee "mean," or "represent," the fact that I have strained the torn ligament again? But the sense of "meaning" involved here seems something like causal indication; the pain "means" a damaged ligament in the same sense in which your nice new suntan means that you spent the weekend on the beach. Prima facie, many bodily sensations don't seem to be evaluable in term of truth or correctness. Or consider moods, like being bored, feeling low and blue, feeling upbeat, and the like. Do they represent anything? Can they be accurate or inaccurate? However, the view that all states of consciousness, including bodily sensations, are representational in nature has recently been gaining in popularity and influence, and we will revisit this issue later (in chapters 9 and 10).

Second, it may be observed that minds, or mental states, are not the only things that exhibit intentionality. Languages, in particular words and sentences, refer to things and have representational contents. The word "London"

refers to London, and the sentence "London is large" refers to, or represents, the fact, or state of affairs, that London is large. A string of zeros and ones in a computer data structure can mean your name and address, and such strings are ultimately electronic states of a physical system. If these physical items and states are capable of reference and content, how can intentionality be considered an exclusive property of mentality?

The following line of reply seems open, however. As some have argued, we might distinguish between *genuine*, or *intrinsic*, intentionality, which our minds and mental states possess, and *as-if*, or *derivative*, intentionality, which we attribute to objects and states that do not have intentionality in their own right.[15] When I say that my computer printer "likes" to work with Windows XP but not with Windows Vista, I am not really saying that my printer has likes and dislikes. It is at best an "as-if" or metaphorical use of language, and no one will take my statement to imply the presence of mentality in the machine. And it seems not implausible to argue that the word "London" refers to London only because language users use the word to refer to London. If we used it to refer to Paris, it would mean Paris, not London. Or if the inscription "London" were not a word in a language, it would just be meaningless scribbles with no referential function. Similarly, the sentence "London is large" represents the state of affairs it represents only because speakers of English use this sentence to represent that state of affairs—for example, in affirming this sentence, they express the belief that London is large. The point, then, is that the intentionality of language is derived from the intentionality of language users and their mental processes. It is the latter that have intrinsic intentionality, intentionality that is not derived from, or borrowed from, anything else. Or so one could argue.[16]

A more direct reply would be this: To the extent that some physical systems can be said to refer to things, represent states of affairs, and deal in meanings, they should be considered as exhibiting mentality, at least one essential form of it. No doubt, as the first reply indicates, analogical or metaphorical uses of intentional idioms abound, but this fact should not blind us to the possibility that physical systems and their states might possess genuine intentionality and hence mentality. After all, it might be argued, we are complex physical systems ourselves, and the physical-biological states of our brains are capable of referring to things and states of affairs external to them and of storing their representations in memory. Of course, it may turn out not to be possible for purely physical states to have such capabilities, but that would only show that they are not capable of mentality. It remains true, the reply goes, that intentionality is at least a sufficient condition for mentality.

A Question

In surveying these candidates for "the mark of the mental," we realize that our notion of the mental is far from unified and monolithic and that it is in fact a cluster of many ideas. Some of the ideas are fairly closely related to one another, but others appear independent of each other. (Why should there be a connection between special epistemic access and nonspatiality?) The diversity and possible lack of unity in our conception of the mental would imply that the class of things and states that we classify as mental may be a varied and heterogeneous lot. It is standardly thought that there are two broad categories of mental phenomena: first, conscious states, in particular sensory or qualitative states (those with "qualia"), like pains and sensings of colors and textures, and, second, intentional states, states with representational contents, like beliefs, desires, and intentions. The former seem to be paradigm cases of states that satisfy the epistemic criteria of the mental, such as direct access and privacy, and the latter are the prime examples of mental states that satisfy the intentionality criterion. An important question that is still open is this: In virtue of *what common property* are both sensory states and intentional states "mental"? What do pains and beliefs have in common in virtue of which they both fall under the single rubric of "mental phenomena"?

There are two approaches that might yield an answer—and a unified conception of mentality. Some have argued that consciousness is fundamental, and that it is presupposed by intentionality—in particular, that all intentional states are either conscious or in principle possible to become conscious.[17] Along the same line, one might urge that only beings with consciousness are capable of having thoughts with content and intentionality. Such a view opens the possibility that all mentality is at bottom anchored in consciousness, and that consciousness is the single foundation of minds.

In direct opposition to this, there is the increasingly influential view, mentioned above, that all consciousness, including phenomenal consciousness, is representational in character. It is held that it is of the essence of conscious states that they represent things to be in a certain way, and that this is no less the case with bodily sensations, like pain, than with perceptual experiences like seeing a green vase on the table. This would mean that all conscious states have representational, or intentional, contents and are "directed upon" the objects and properties represented. Representationalism about consciousness, therefore, leads to the view that intentionality is the single mark characterizing all mentality. Thus, one potential bonus from consciousness representationalism could be a satisfying unified concept of minds and mentality.

For Further Reading

Readers interested in philosophical issues of cognitive science may explore Andy Clark, *Mindware: An Introduction to the Philosophy of Cognitive Science*; Barbara von Eckardt, *What Is Cognitive Science?*; Robert M. Harnish, *Minds, Brains, Computers: An Historical Introduction to the Foundations of Cognitive Science*. Also useful are two anthologies: *Minds, Brains, and Computers*, edited by Denise Dellarosa Cummins and Robert Cummins; *Readings in Philosophy and Cognitive Science*, edited by Alvin Goldman.

The Oxford Handbook of Philosophy of Mind, edited by Brian McLaughlin, Ansgar Beckermann, and Sven Walter, is a comprehensive and highly useful reference work. The following general encyclopedias of philosophy feature many fine articles (some with extensive bibliographies) on topics in philosophy of mind and related fields: *Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu); *Macmillan Encyclopedia of Philosophy*, second edition, edited by Donald Borchert; and *Routledge Encyclopedia of Philosophy*, edited by Edward Craig. The "Mind & Cognitive Science" section of *Philosophy Compass* (www.blackwell-compass.com) includes many fine up-to-date surveys of current research on a variety of topics in philosophy of mind. *The Internet Encyclopedia of Philosophy* (www.iep.utm.edu) has many helpful entries in its "Mind & Cognitive Science" section. In general, however, readers should exercise proper caution when consulting Web resources.

There are many good general anthologies on philosophy of mind. To mention a sample: *The Philosophy of Mind*, edited by Brian Beakley and Peter Ludlow; *Philosophy of Mind: Classical and Contemporary Readings*, edited by David J. Chalmers; *Problems in Mind*, edited by Jack S. Crumley II; *Philosophy of Mind: A Guide and Anthology*, edited by John Heil; *Mind and Cognition: An Anthology*, third edition, edited by William G. Lycan and Jesse Prinz; *Philosophy of Mind: Contemporary Readings*, edited by Timothy O'Connor and David Robb.

Notes

1. For details see Brian McLaughlin and Karen Bennett, "Supervenience."

2. Sometimes this version of supervenience is formulated as follows: "Any minimal physical duplicate of this world is a duplicate *simpliciter* of this world." See, for example, Frank Jackson, "Finding the Mind in the Natural World." The point of the qualifier "minimal" is to exclude the following kind of situation: Consider a world that is like ours in all physical respects but in addition contains ectoplasms and immaterial spirits. (We are assuming these

things do not exist in the actual world.) There is a sense in which this world and our world are physically alike, but they are clearly not alike overall. A case like this is ruled out by the qualifier "minimal" because this strange world is not a minimal physical duplicate of our world.

3. On characterizing physicalism, see Alyssa Ney, "Defining Physicalism."

4. Also called "ontological physicalism."

5. Nonreductive physicalism, as a form of physicalism, also includes mind-body supervenience; property dualism as such is not committed to supervenience. In fact, Cartesian substance dualism entails property dualism.

6. We should keep in mind the possibility that these philosophers who accept supervenience but reject reducibility are just mistaken.

7. For more on token and type physicalism, see Jaegwon Kim, "The Very Idea of Token Physicalism."

8. Why they are called "intentional" states is not simple to explain or motivate; it is best taken simply as part of philosophical terminology. If you insist on an explanation, the following might help: These states, in virtue of their contents, are *representational* states; the belief that snow is white represents the world as being a certain way—more specifically, it represents the state of affairs of snow being white. Traditionally, the term "intentionality" has been used to refer to this sort of representational character of mental states. More to follow on intentionality below.

9. René Descartes, *Meditations on First Philosophy*, Meditation II.

10. Later in the book (chapter 8) you will encounter another sense of "transparency" applied to perceptual experiences.

11. It is worth noting that many psychologists and cognitive scientists take a dim view of the claim that we have specially privileged access to the contents of our minds. See, for example, Richard Nisbett and Timothy Wilson, "Telling More Than We Can Know," and Alison Gopnik, "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality."

12. Franz Brentano, *Psychology from an Empirical Standpoint*, p. 88.

13. Ludwig Wittgenstein, *Philosophical Investigations*, p. 177.

14. This example is taken from Ned Block.

15. See, for example, John Searle, *Intentionality* and *The Rediscovery of the Mind*.

16. This point has been disputed. Other possible positions are these: First, one might hold that linguistic intentionality is in fact prior to mental intentionality, the latter being derivative from the former (Wilfrid Sellars); second, we might claim that the two types of intentionality are distinct but interdependent, neither being prior to the other and neither being derivable from the

other (Donald Davidson); and third, some have argued that the very distinction between "intrinsic" and "derivative" intentionality is bogus and incoherent (Daniel Dennett).

17. John Searle is a well-known advocate of this claim; see his *The Rediscovery of Mind*, chapter 7. See also Galen Strawson, "Real Intentionality 3: Why Intentionality Entails Consciousness."

CHAPTER 2

Mind as Immaterial Substance

Descartes's Dualism

What is it for something to "have a mind," or "have mentality"? When the ancients reflected on the contrast between us and mindless creatures, they sometimes described the difference in terms of having a "soul." For example, according to Plato, each of us has a soul that is simple, divine, and immutable, unlike our bodies, which are composite and perishable. In fact, before we were born into this world, our souls preexisted in a pure, disembodied state, and on Plato's doctrine of recollection, what we call "learning" is merely a process of recollecting what we already knew in our prenatal existence as pure souls. Bodies are merely vehicles of our existence in this earthly world, a transitory stage in our soul's eternal journey. The idea, then, is that because each of us has a soul, we are the kind of conscious, intelligent, and rational creatures that we are. Strictly speaking, we do not really "have" souls, since we are literally *identical with* our souls—that is, each of us *is* a soul. My soul is the thing that I am. Each of us "has a mind," therefore, because each of us *is* a mind.

For most of us, Plato's story is probably a bit too speculative, too fantastical, to take seriously as a real possibility. However, many of us seem to have internalized a kind of mind-body dualism according to which, although each of us has a body that is fully material, we also have a mental or spiritual dimension that no "mere" material things can have. When we see the term "material," we are apt to think "not mental" or "not spiritual," and when we see the term "mental," we tend to think "not material" or "not physical." This may not amount to a clearly delineated point of view, but it seems fair to say that some

such dualism of the mental and the material is entrenched in our ordinary thinking, and that dualism is a kind of "folk" theory of our nature as creatures with minds.

But folk dualism often goes beyond a mere duality of mental and physical properties, activities, and processes. It is part of folklore in many cultures and of most established religions that, as Plato claimed, each of us has a soul, or spirit, that survives bodily death and decay, and that we are really our souls, not our bodies, in that when our bodies die we continue to exist in virtue of the fact that our souls continue to exist. Your soul defines your identity as an individual person; as long as it exists—and only so long as it exists—you exist. And it is our souls in which our mentality inheres; thoughts, consciousness, rational will, and other mental acts, functions, and capacities belong to souls, not to material bodies. Ultimately, to have a mind, or to be a creature with mentality, is to have a soul.

In this chapter, we examine a theory of mind, due to the seventeenth-century French philosopher René Descartes, which develops a view of this kind. One caveat before we begin: Our goal here is not so much a scholarly exegesis of Descartes as it is an examination of a point of view closely associated with him. As with other great philosophers, the interpretation of what Descartes "really" said, or meant to say, continues to be controversial. For this reason, the dualist view of mind we will discuss is better regarded as Cartesian rather than as the historical Descartes's.

Descartes's Interactionist Substance Dualism

The dualist view of persons that Descartes defended is a form of substance dualism (sometimes called substantial, or substantival, dualism). Substance dualism is the thesis that there are substances of two fundamentally distinct kinds in this world, namely, minds and bodies—or mental stuff and material stuff—and that a human person is a composite entity consisting of a mind and a body, each of which is an entity in its own right. Dualism of this form contrasts with monism, according to which all things in the world are substances of one kind. We later encounter various forms of material monism that hold that our world is fundamentally material, consisting only of bits of matter and complex structures made up of bits of matter, all behaving in accordance with physical laws. This is materialism, or physicalism. (The terms "materialism" and "physicalism" are often used interchangeably, although there are subtle differences: We can think of physicalism as a contemporary successor to materialism—

materialism informed by modern physics.) There is also a mental version of monism, unhelpfully called idealism. This is the view that minds, or mental items at any rate ("ideas"), constitute the fundamental reality of the world, and that material things are mere "constructs" out of thoughts and mental experiences. This form of monism has not been very much in evidence for some time, though there are reputable philosophers who still defend it.[1] We will not be further concerned with mental monism in this book.

So substance dualism maintains that minds and bodies are two different sorts of substance. But what is a substance? Traditionally, two ideas have been closely associated with the concept of a substance. First, a substance is something in which properties "inhere"; that is, it is what has, or instantiates, properties.[2] Consider this celadon vase on my table. It is something that has properties, like weight, shape, color, and volume; it is also fragile and elegant. But a substance is not in turn something that other things can exemplify or instantiate; nothing can have, or instantiate, the vase as a property. Linguistically, this idea is sometimes expressed by saying that a substance is the subject of predication, something to which we can attribute predicates like "blue," "weighs a pound," and "fragile," while it cannot in turn be predicated of anything else.

Second, and this is more important for us, a substance is thought to be something that has the capacity for independent existence. Descartes himself wrote, "The notion of a substance is just this—that it can exist by itself, that is without the aid of any other substance."[3] What does this mean? Consider the vase and the pencil holder to its right. Either can exist without the other existing; we can conceive the vase as existing without the pencil holder existing, and vice versa. In fact, we can, it seems, conceive of a world in which only the vase (with all its constituent parts) exists and nothing else, and a world in which only the pencil holder exists and nothing else. It is in this sense that a substance is capable of independent existence. This means that if my mind is a substance, it can exist without any body existing, or any other mind existing. Consider the vase again: There is an intuitively intelligible sense in which its color and shape cannot exist apart from the vase, whereas the vase is something that exists in its own right. (The color and shape would be "modes" that belong to the vase.) The same seems to hold when we compare the vase and its surface. Surfaces are "dependent entities," as some would say; their existence depends on the existence of the objects of which they are surfaces, whereas an object could exist without the particular surface it happens to have at a given time. As was noted, there is a possible world of which the vase is the sole inhabitant. Compare the evidently absurd claim that there is a possible world in which the surface of the vase exists but nothing else; in fact,

there is no possible world in which only surfaces exist and nothing else. For surfaces to exist they must be surfaces of some objects—existing objects.[4]

Thus, the thesis that minds are substances implies that minds are objects, or things, in their own right; in this respect, they are like material objects—it's only that, on Descartes's view, they are immaterial objects. They have properties and engage in activities of various sorts, like thinking, sensing, judging, and willing. Most important, they are capable of independent existence, and this means that there is a possible world in which only minds exist and nothing else—in particular, no material bodies. So my mind, as a substance, can exist apart from my body, and so of course could your mind even if your body perished.

Let us put down the major tenets of Cartesian substance dualism:

1. There are substances of two fundamentally different kinds in the world, mental substances and material substances—or minds and bodies. The essential nature of a mind is to think, be conscious, and engage in other mental activities; the essence of a body is to have spatial extensions (a bulk) and be located in space.
2. A human person is a composite being (a "union," as Descartes called it) of a mind and a body.
3. Minds are diverse from bodies; no mind is identical with a body.

What distinguishes Descartes's philosophy of mind from the positions of many of his contemporaries, including Leibniz, Malebranche, and Spinoza, is his eminently commonsensical belief that minds and bodies are in causal interaction with each other. When we perform a voluntary action, the mind causes the body to move in appropriate ways, as when my desire for water causes my hand to reach for a glass of water. In perception, causation works in the opposite direction: When we see a tree, the tree causes in us a visual experience as of a tree. That is the difference between seeing a tree and merely imagining or hallucinating one. Thus, we have the following thesis of mind-body causal interaction:

4. Minds and bodies causally influence each other. Some mental phenomena are causes of physical phenomena and vice versa.

The only way we can influence the objects and events around us, as far as we know, is first to move our limbs or vocal cords in appropriate ways and thereby start a chain of events culminating in the effects we desire—like opening a window, retrieving a hat from the roof, or starting a war. But as we will see, it is this most plausible thesis of mind-body causal interaction that

brought down Cartesian dualism. The question was not whether the interactionist thesis was in itself acceptable; rather, the main question was whether it was compatible with the radical dualism of minds and bodies—that is, whether minds and bodies, sundered apart by the dualist theses (1) and (3), could be brought together in causal interaction as claimed in (4).

Why Minds and Bodies Are Distinct: Some Arguments

Before we consider the supposed difficulties for Descartes's interactionist dualism, let us first consider some arguments that apparently favor the dualist thesis that minds are distinct from bodies. Most of the arguments we will consider are Cartesian—some of them perhaps only vaguely so—in the sense that they can be traced one way or another to Descartes's *Second* and *Sixth Meditations* and that all are at least Cartesian in spirit. It is not claimed, however, that these are in fact the arguments that Descartes offered or that they were among the considerations that moved Descartes to advocate substance dualism. You might want to know first of all why anyone would think of minds as substances—why we should countenance minds as objects or things in addition to people and creatures with mentality. As we will see, some of the arguments do address this issue, though not directly.

At the outset of his *Second Meditation*, Descartes offers his famous "cogito" argument. As every student of philosophy knows, the argument goes "I think, therefore I exist." This inference convinces him that he can be absolutely certain about his own existence; his existence is one perfectly indubitable bit of knowledge he has, or so he is led to think. Now that he knows he exists, he wonders what kind of thing he is, asking, "But what then am I?" Good question! Knowing that you exist is not to know very much; it has little content. So what kind of being is Descartes? He answers: "A thinking thing" ("sum res cogitans"). How does he know that? Because he has proved his existence from the premise that he thinks; it is through his knowledge of himself as a thinker that he knows that he exists. To get on with his dualist arguments we will grant him the proposition that he is a thinking thing, namely a mind. The main remaining issue for him, and for us, is the question whether the thinking thing can be his body—that is, why we should not take his body, perhaps his brain, as the thing that does the thinking.

We first consider three arguments based on epistemological considerations. The simplest—perhaps a bit simplistic—argument of this form would be something like this:

Argument 1

I am such that my existence cannot be doubted.
My body is not such that its existence cannot be doubted.
Therefore, I am not identical with my body.
Therefore, the thinking thing that I am, that is, my mind, is not identical with my body.

This argument is based on the apparent asymmetry between knowledge of one's own existence and knowledge of one's body's existence: While I cannot doubt that I exist, I can doubt that my body exists. We could also put the point this way: As the cogito argument shows, I can be absolutely certain that I exist, but my knowledge that my body exists, or that I have a body, does not enjoy the same degree of certainty. I must make observations to know that I have a body, and such observations could go astray. We leave it to the reader to evaluate this argument.

According to Descartes, I am a "thinking thing." What does this mean? He says that a thinking thing is "a thing that doubts, understands, affirms, denies, is willing, is unwilling, and also imagines and has sensory perceptions."[5] For Descartes, then, "thinking" is a generic term, roughly meaning "mental activity," and specific mental states and activities, like believing, doubting, affirming, reasoning, sensing a color, hearing a sound, experiencing a pain, and the rest fall under the broad rubric of thinking. In Descartes's own terms, thinking is the general essence of minds, and the specific kinds of mental activities and states are its various "modes."

Our second epistemological argument exploits another related difference between our knowledge of our own minds and our knowledge of our bodies.

Argument 2

My mind is transparent to me—that is, nothing can be in my mind without my knowing that it is there.
My body is not transparent to me in the same way.
Therefore, my mind is not identical with my body.

As stated, the first premise is quite strong and likely not to be entirely true. Most of us would be prepared to acknowledge that at least some of our beliefs, desires, and emotions are beyond our cognitive reach—that is, that there are "unconscious" or "subconscious" mental states, like suppressed beliefs and

desires, angers and resentments, of which we are unaware. This, however, doesn't seem like a big problem: The premise can be stated in a weaker form, to claim only that my mind is transparent at least with respect to *some* of the events that occur in it. This weaker premise suffices as long as we understand the second premise as asserting that *none* of my bodily events have this transparent character. To find out any fact about my body, I must make observations and sometimes make inferences from the evidence gained through observations. Often a third party—my physician or dentist—is in a better position to know the conditions of my body.

We now consider our last epistemological argument for substance dualism:

Argument 3

Each mind is such that there is a unique subject who has direct access to its contents.
No material body has a specially privileged knower—knowledge of material things is in principle public and intersubjective.
Therefore, minds are not identical with material bodies.

We are said to know something "directly" when the knowledge is not based on evidence, or inferred from other things we know. When knowledge is direct, like my knowledge of my toothache, it makes no sense to ask, "How do you know?" The present argument exploits this difference between knowledge of minds and knowledge of bodies: For each mind, there is a unique person who is in a privileged epistemic position, whereas this is not the case with bodies. It is in this sense that knowledge of our own minds is said to be "subjective." In contrast, knowledge of bodies is said to be "objective"—different observers can in principle have equal access to such knowledge. Thus, the present argument can be called the argument from the subjectivity of minds.

What should we think of these arguments? We will not formulate and develop specific objections and difficulties, or discuss how the dualist might respond; that is left to the reader. But one observation is in order: It is widely believed that there is a problem with using epistemic (or more broadly, "intentional") properties to differentiate things. To show that $X \neq Y$, it is necessary and sufficient to come up with a single property P such that X has P but Y lacks it, or Y has P but X lacks it. Such a property P can be called a *differential property* for X and Y. The question, then, is whether epistemic properties, like being known with certainty (or an intentional property like being believed to be such and such), can be used as a differential property. Consider

the property of being known to the police to be the hit-and-run driver. The man who sped away in a black SUV is known to the police to be the hit-and-run driver. The man who drove away in a black SUV is identical with my neighbor, and yet my neighbor is not known to the police to be the hit-and-run driver (or else the police would have him in custody already). The epistemic properties invoked in the three arguments are not the same—or exactly of the same sort—as the one just used. It is fair to say that the last of the arguments presented above, the argument from subjectivity, seems the most compelling, and anyone wishing to reject it should have good reasons.

We now turn to metaphysical arguments, which instead of appealing to epistemic differences between minds and bodies attempt to invoke real metaphysical differences between them. Throughout the *Second* and *Sixth Meditations*, there are constant references to the essence of mind as thinking and the essence of body as being extended in space. By extension in space Descartes means three-dimensional extension, that is, bulk. Surfaces or geometric lines do not count as material substances; only things that have a bulk count as such. A simple argument could be formulated in terms of essences or essential natures, like this:

Argument 4

My essential nature is to be a thinking thing.
My body's essential nature is to be an extended thing in space.
My essential nature does not include being an extended thing in space.
Therefore, I am not identical with my body. And since I am a thinking thing (namely a mind), my mind is not identical with my body.

How could the first and third premises be defended? Perhaps a Cartesian dualist could make two points in defense of the first premise. First, as the "cogito" argument shows, I know that I exist only insofar as I am a thinking thing, and this means that my existence is inseparably tied to the fact that I am a thinking thing. Second, an essential nature of something is a property without which the thing cannot exist; when something loses its essential nature, that is when it ceases to exist. Precisely in this sense, being a thinking thing is my essential nature; when I cease to be a thinking thing, that is, a being with a capacity for thought and consciousness, that is when I cease to be, and so long as I am a thinking thing, I exist. On the other hand, I can conceive of myself as existing without a body; there is no inherent incoherence, or contradiction, in the idea of my disembodied existence, whereas it seems mani-

festly incoherent to think of myself as existing without a capacity to think and have conscious experience. Hence, being an extended object in space is not part of my essential nature.

What should we think of this argument? Some will question how the third line of the argument might be established, pointing out that all Descartes shows is that our disembodied existence is conceivable, or imaginable. But from the fact that something is *conceivable*, however clearly and vividly, it does not follow that it is *really possible*. A body moving at a speed exceeding the speed of light is conceivable, but we know it is not possible.[6] Or consider this: We seem to be able to conceive how Goldbach's conjecture, the proposition that every even number greater than two is the sum of two prime numbers, might turn out to be true, and also to conceive how it might turn out to be false. But Goldbach's conjecture, being a mathematical proposition, is necessarily true if true, and necessarily false if false. So it cannot be both possibly true and possibly false. (To the reader: Why?) But if conceivability entails possibility, it would have to be possibly true and possibly false. This issue about conceivability and real possibility has led to an extended series of debates too complex to enter into here.[7] It is a live current issue in modal metaphysics and epistemology. We should note, though, that unless we use reflective and carefully scrutinized conceivability as a guide to possibility, it is difficult to know what other resources we can call on when we try to determine what is possible and what is not, what is necessarily the case and what is only contingently so, and other such modal questions.

Let us say that something is "essentially" or "necessarily" F, where "F" denotes a property, just in case whenever or wherever it exists (or in any possible world in which it exists), it is F. In this sense, we are presumably essentially persons, but not essentially students or teachers; for we cannot continue to exist while ceasing to be persons, whereas we could cease to be students, or teachers, without ceasing to exist. In the terminology of the preceding paragraph, for something to have property F essentially or necessarily is to have F as part of its essential nature. Consider, then, the following argument:

Argument 5

If anything is material, it is essentially material.
However, I am possibly immaterial—that is, there is a possible world in which I exist without a body.
Hence, I am not essentially material.
Hence, it follows (with the first premise) that I am not material.

This is an interesting argument. There seems to be a lot to be said for the first premise. Take something material, say, a bronze bust of Beethoven: This object could perhaps exist without being a bust of Beethoven—it could have been fashioned into a bust of Brahms. In fact, it could exist without being a bust of anyone; it could be melted down and made into a doorstop. If transmutation of matter were possible (surely this is not something a priori impossible), it could even exist without being bronze. But could this statue exist without being a material thing? The answer seems a clear no. If anything is a material object, being material is part of its essential nature; it cannot exist without being a material thing. So it appears that the acceptability of the argument depends crucially on the acceptability of the second premise. Is it possible that I exist without a body? That surely is conceivable, Descartes would insist. But again, is something possible just because it is conceivable? Can we say more about the possibility of our disembodied existence?

Consider the bronze bust again. There is here a piece of sculpture and a quantity of bronze. Is the sculpture the very same thing as the bronze? Many philosophers would say no: Although the two share many properties in common (such as weight, density, and location), they differ at least in one respect, namely, their persistence condition. If the bust is melted down and shaped into a cube, the bust is gone but the bronze continues to exist. According to the next dualist argument, my body and I differ in a somewhat similar way.

Argument 6

Suppose I am identical with this body of mine.
In 2001 this body did not exist.
Hence, from the first premise, it follows that I did not exist in 2001.
But I existed in 2001.
Hence, a contradiction, and the supposition must be false.
Hence, I am not identical with my body.

In 2001 this body did not exist because all the molecules making up a human body are completely cycled out every six or seven years. When all the molecular constituents of a material thing are replaced, we have a new material thing. The body that I now have shares no constituents with the body I had in 2001. The person that I am, however, persists through changes of material constituents. So even if I have to have some material body or other to exist, I do not have to have any particular body. But if I am identical with a

body, I must be identical with some particular body and when this body goes, so go I. That is the argument. (This probably was not one of Descartes's actual arguments.)

An initial response to this argument could run as follows: When I say I am identical with this body of mine, I do not mean that I am identical with the "time slice"—that is, a temporal cross section—of my body at this instant. What I mean is that I am identical with the temporally elongated "worm" of a three-dimensional organism that came into existence at my birth and will cease to exist when my biological death occurs. This four-dimensional object—a three-dimensional object stretched along the temporal dimension—has different material constituents at different times, but it is a clearly delineated system with a substantival unity and integrity. It is this material structure with a history with which I claim I am identical. (To the reader: How might a Cartesian dualist reformulate the argument in answer to this objection?)

Another reply, related to the first, could go as follows: My body is not a mere assemblage or structure made up of material particles; rather, it is a biological organism, a human animal. And the persistence condition appropriate to mere material things is not necessarily appropriate for animals. In fact, animals can retain their identities even though the matter constituting them changes over time (this may well be true of all living things, including plants), just as in the case of persons. The criterion of identity over time for animals (however it is to be spelled out in detail) is the one that should be applied to human bodies.[8] Does the substance dualist have a reply to this? I believe an answer may be implicit in the next argument we consider.

Tully is the same person as Cicero. There is one person here, not two. Can there be a time at which Tully exists but not Cicero? Obviously not—that is no more possible than for Tully to be at a place where Cicero is not. Given that Cicero = Tully in this world, is there a possible world in which Cicero is not identical with Tully? That is, given that Cicero is Tully, is it *possible* that Cicero is not Tully? Suppose there is a possible world in which Cicero ≠ Tully; call it W. Since Cicero ≠ Tully in W, there must be some property, F, such that, in W, Cicero has it but Tully does not. Let's say that F is the property of being tall. So in W, Cicero is tall but Tully isn't. But how is that possible? Here in this world is a single person, called both Cicero and Tully. How is it possible for this one person to be tall and at the same time not tall in world W? That surely is an impossibility, and world W is not a possible world. In fact, there is no possible world in which Cicero ≠ Tully. We therefore have the following principle ("NI" for "necessity of identities"):

(NI) If X = Y, then necessarily X = Y—that is, if X = Y in this world, X = Y in every possible world.

(NI) is special in that in general it is not the case that if a proposition is true, it is necessarily true. For example, I am standing; from this it does not follow that necessarily I am standing, for I could be sitting.

Given the principle (NI), we can formulate another dualist argument:

Argument 7

Suppose I am identical with this body of mine.
Then, by (NI), I am necessarily identical with this body—that is, I am identical with it in every possible world.
But that is false, for (a) in some possible worlds I could be disembodied and have no body; or at least (b) I could have a *different* body in another possible world.
So it is false that I am identical with this body in every possible world, and this contradicts the second line.
Therefore, I am not identical with my body.

The principle (NI) is considered unexceptionable. So if there is a vulnerability in this argument, it would have to be the third line; to criticize this premise effectively, we would have to eliminate both (a) and (b) as possibilities. As we have seen, (a) is vulnerable to criticism; however, (b) may be less so. John Locke's well-known story of the prince and the cobbler can be taken as supporting (b); Locke writes:

> Should the soul of a prince, carrying with it the consciousness of the prince's past life, enter and inform the body of a cobbler, as soon as deserted by his own soul, everyone sees he would be the same person with the prince, accountable only for the prince's action. . . . Had I the same consciousness that I saw the ark and Noah's flood, as that I saw an overflowing of the Thames last winter, or as that I write now, I could no more doubt that I who write this now, that saw the Thames overflowed last winter, that viewed the flood at the general deluge, was the same *self* . . . than that I who write this am the same *myself* now whilst I write . . . that I was yesterday.[10]

For Locke, then, consciousness, not body, defines a person, a self; the continuity of my consciousness determines my persistence as a person. What body I

have, or whether I have a body at all, is immaterial. To defeat this dualist argument, therefore, we must show that Locke's story of the prince and the cobbler is an impossibility—it isn't something that could happen. This will require some ingenuity and creative thinking.

The leading idea driving all of these metaphysical arguments is the thought that although I may be a composite being consisting of a mind and a body, my relation to my mind is more intimate and essential than my relation to my body and that I am "really" my mind and could not exist apart from it, while it is a contingent fact that I have the body that I happen to have. Descartes's interest in defending minds as immaterial substances was apparently motivated in part by his desire to allow for the possibility of survival after bodily death.[11] Most established religions have a story to tell about the afterlife, and the conceptions of an afterlife in some of them seem to require, or at least allow, the possibility of our existence without a body. But all that is a wish list; it does not make the possibility of our disembodied existence a real one (Descartes was under no such illusion). The arguments we have looked at must earn their plausibility on their own merits, not from the allure of their conclusions.

We will now consider our final metaphysical argument for substance dualism. As we will see, this argument is rather difficult to articulate clearly, but it enjoys the allegiance of some well-known and well-respected philosophers, so it is worth a serious look. The skeletal structure of the argument can be set out like this:

Argument 8

Thoughts and consciousness exist.
Hence, there must be objects, or substances, to which thoughts and consciousness occur—that is, things that think and are conscious.
Thoughts and consciousness cannot occur to material things—they cannot be states of material objects, like the brain.
Hence, thoughts and consciousness must occur to immaterial things, like Cartesian mental substances.
Hence, mental substances exist and they are the things that think and are conscious, and bear other mental properties.

Some would question the move from the first to the second line—the assumption that thoughts and consciousness, and, more generally, states and properties, require "bearers," things to which they occur, or in which they inhere; this, however, is a general metaphysical issue and it will be tedious and

out of place to pursue it here. Moreover, the crucial premise is staring us in the face—it is the third line, the proposition that material things, like the human brain, are unfit to serve as bearers of thoughts and consciousness. Think about numbers, like three and fifteen: Numbers aren't the sort of thing that can have colors like blue or red, or occupy a location in space, or be transparent or opaque. Or think about events, like earthquakes or wildfires. They can be sudden, severe, and destructive; but events aren't the sort of thing that can be soluble in water, divisible by four, or weigh ten tons. The claim then is that there is an essential incongruity between mental states, like thoughts and consciousness, on one hand and material things on the other, so that the former cannot inhere in, or occur to, the latter, just as weight and color cannot inhere in numbers. If our thoughts and consciousness cannot occur to anything material, including our brains, then they must occur to immaterial things, or Cartesian minds. Only immaterial things can be conscious and have thoughts. Since we are conscious and have thoughts, we must be immaterial minds.

But why can't consciousness, thoughts, and other mental states occur to material things? It is often thought that Leibniz was first to give an argument, or at least hint at one, why that must be so:

> It must be confessed, moreover, that *perception*, and that which depends on it, *are inexplicable by mechanical causes*, that is by figures and motions. Supposing there were a machine so constructed as to think, feel and have perception, we could conceive of it as enlarged and yet preserving the same proportions, so that we might enter it as into a mill. And this granted, we should only find, on visiting it, pieces which push one against another, but never anything by which to explain a perception. This must be sought for, therefore, in the simple substance and not in the composite or in the machine.[12]

Leibniz appears to be saying that a material thing is at bottom a mechanical system in which the parts causally interact with one another ("pieces pushing one against another"), and it is not possible to see anything in this picture that would account for the presence of thought or consciousness. This is not altered when a more sophisticated modern picture of what goes on in a complex biological system, like a human brain, replaces Leibniz's mill: What we have is still a large assemblage of microscopic material things, molecules and atoms and particles, interacting with one another in accordance with laws of physics and chemistry, producing further scenes of such interactions. Nowhere in this

picture do we see a thought or perception or consciousness; molecules jostling and bumping against one another is all the action that is taking place. Again, if this picture looks unsophisticated, replace it with the most sophisticated scientific picture you know, and see if that invalidates Leibniz's point.

Is this all one can say in defense of the Leibnizian proposition that material systems are just the wrong kind of thing to bear thoughts and other mental states? It might be helpful to consider what some philosophers have said to defend this proposition. Alvin Plantinga, referring to the Leibniz paragraph above, writes:

> Leibniz's claim is that thinking can't arise by virtue of physical interaction among objects or parts of objects. According to current science, electrons and quarks are simple, without parts. Presumably neither can think—neither can adopt propositional attitudes; neither can believe, doubt, hope, want, or fear. But then a proton composed of quarks won't be able to think either, at least by way of physical relations between its component quarks, and the same will go for an atom composed of protons and electrons, and a molecule composed of atoms, a cell composed of molecules, and an organ (e.g., a brain) composed of cells. If electrons and quarks can't think, we won't find anything composed of them that *can* think by way of the physical interaction of its parts.[13]

Does this reading of Leibniz shed new light on his argument and make it seem more plausible? It is something to ponder. Some, for example the emergentists, will argue that thoughts and consciousness arise in material systems when they reach higher levels of organizational complexity, and that from the fact that the constituent parts of a system lack a certain property it does not follow that the system itself must lack that property.

Another philosopher, John Foster, who holds the view that subjects of mentality must be "wholly nonphysical," argues:

> If something is just an ordinary material object, whose essential nature is purely physical, there seems to be no way of understanding how it could be [the subject] of mentality. . . . If something is merely a material object, any understanding of how it is equipped to be a mental subject will presumably have to be achieved by focusing on its physical nature. But focusing on an object's physical nature will only reveal how it is equipped to be in states or engage in activities which are directly to do with its possession of

> that nature—with its condition as a physical thing. . . . Focusing on the physical nature of an object simply offers no clue as to how it can be the basic subject of the kinds of mentality which the dualist postulates.[14]

Perhaps some readers will find these quotations helpful and clarifying; others may not. In any case, one question we should ask at this point is this: Is it any easier to understand how thoughts and consciousness can arise in an immaterial substance, especially if, as Leibniz and many other dualists urge, such a substance is an absolute "simple" with no constituent parts? How could immaterial minds, without structure and outside physical space, possess beliefs and desires directed at things in the physical world? How could our rich and complex mental life inhere in something that has no parts and hence no structure? Isn't the proposal recommended by Leibniz, and by Plantinga and Foster, merely a solution by stipulation? What do we know about mental substances that can help us understand how they could be the bearers of consciousness and perception and thought? Understanding how mentality can arise in something immaterial may be no easier than understanding how it could arise in a material system; in fact, it might turn out to be more difficult.

As was mentioned above, it is not easy to make clear the thoughts that lie behind Argument 8, in particular its crucial third line. However, this is an intriguing and influential line of dualist thinking, and readers are urged to reflect on it.[15]

Princess Elisabeth Against Descartes

As will be recalled, the fourth component of Descartes's dualism is the thesis that minds and bodies causally influence each other. In voluntary action, the mind's volition causes our limbs to move; in perception, physical stimuli impinging on sensory receptors cause perceptual experiences in the mind. This view is not only commonsensical but also absolutely essential to our conception of ourselves as agents and cognizers: Unless our minds, in virtue of having certain desires, beliefs, and intentions, are able to cause our bodies to move in appropriate ways, how could human agency be possible? How could we be agents who act and take responsibility for our actions? If objects and events in the physical world do not cause us to have perceptual experiences and beliefs, how could we have any knowledge of what is happening around us? How could we know that we are holding a tomato in our hand, that we are coming up on a stop sign, or that a large bear is approaching from our left?

Descartes has something to say about how mental causation works. In the *Sixth Meditation*, he writes:

> The mind is not immediately affected by all parts of the body, but only by the brain, or perhaps just by one small part of the brain. . . . Every time this part of the brain is in a given state, it presents the same signals to the mind, even though the other parts of the body may be in a different condition at the time. . . . For example, when the nerves in the foot are set in motion in a violent and unusual manner, this motion, by way of the spinal cord, reaches the inner parts of the brain, and there gives the mind its signal for having a certain sensation, namely the sensation of a pain as occurring in the foot. This stimulates the mind to do its best to get rid of the cause of the pain, which it takes to be harmful to the foot.[16]

In *The Passions of the Soul*, Descartes identifies the pineal gland as the "seat of the soul," the locus of direct mind-body interaction. This gland, Descartes maintains, can be moved directly by the soul, thereby moving the "animal spirits" (bodily fluids in the nerves), which then transmit causal influence to appropriate parts of the body:

> And the activity of the soul consists entirely in the fact that simply by willing something it brings it about that the little gland to which it is closely joined moves in the manner required to produce the effect corresponding to this desire.[17]

In the case of physical-to-mental causation, this process is reversed: Disturbances in the animal spirits surrounding the pineal gland make the gland move, which in turn causes the mind to experience appropriate sensations and perceptions. For Descartes, then, each of us as an embodied human person is a "union" or "intermingling" of a mind and a body in direct causal interaction.

In what must be one of the most celebrated letters in the history of philosophy, Princess Elisabeth of Bohemia, an immensely astute pupil of Descartes's, wrote to him in May 1643, challenging him to explain

> how the mind of a human being, being only a thinking substance, can determine the bodily spirits in producing bodily actions. For it appears that all determination of movement is produced by the pushing of the thing being moved, by the manner in which it is pushed by that which moves it,

> or else by the qualification and figure of the surface of the latter. Contact is required for the first two conditions, and extension for the third. [But] you entirely exclude the latter from the notion you have of the soul, and the former seems incompatible with an immaterial thing.[18]

(For "determine," read "cause"; for "bodily spirits," read "fluids in the nerves and muscles.") Elisabeth's demand is clearly understandable. First, see what Descartes has said about bodies and their motion in the *Second Meditation*:

> By a body I understand whatever has determinate shape and a definable location and can occupy a space in such a way as to exclude any other body; it can be perceived by touch, sight, hearing, taste or smell, and can be moved in various ways, not by itself but by whatever else comes into contact with it.[19]

For Descartes, minds are immaterial; that is, minds have no spatial extension and are not located in physical space. If bodies can be moved only by contact, how could an unextended mind, which is not even in space, come into contact with an extended material thing, even the finest and lightest particles in animal spirits, thereby causing it to move? This seems like a perfectly reasonable question.

In modern terminology we can put Elisabeth's challenge as follows: For anything to cause a physical object to move, or cause any change in one, there must be a flow of energy, or transfer of momentum, from the cause to the physical object. But how could there be an energy flow from an immaterial mind to a material thing? What kind of energy could it be? How could anything "flow" from something *outside space* to something *in space*? If an object is going to impart momentum to another, it must have mass and velocity. But how could an unextended mind outside physical space have either mass or velocity? The question does not concern the intrinsic plausibility of Descartes's thesis of mind-body interaction; the question is whether this commonsensical interactionist thesis is tenable within Descartes's dualist ontology of nonspatial immaterial minds and material things in the space-time world.

Descartes responded to Elisabeth in a letter written in the same month:

> I observe that there are in us certain primitive notions which are, as it were the originals on the pattern of which we form all of other thoughts, . . . as regards the mind and body together, we have only the primitive notion of their union, on which depends our notion of the mind's power to move

> the body, and the body's power to act on the mind and cause sensations and passions.[20]

Descartes is defending the position that the idea of mind-body union is a "primitive" notion—a fundamental notion that is intelligible in its own right and cannot be explained in terms of other more basic notions—and that the idea of mind-body causation depends on that of mind-body union. What does this mean? Although on Descartes's view, minds and bodies seem on an equal footing causally, there is an important asymmetry between them: My mind can exercise its causal powers—on other minds as well as on bodies around me—only by first causally influencing my own body, and nothing can causally affect my mind except through its causal influence on my body. But my body is different: It can causally interact with other bodies quite independently of my mind. My body—or my pineal gland—is the necessary causal conduit between my mind and the rest of the world; in a sense, my mind is causally isolated from the world by being united with my body. To put it another way, my body is the enabler of my mind's causal powers; it is by being united with my body that my mind can exercise its causal powers in the world—on other minds as well as on other bodies. Looked at this way, the idea of mind-body union does seem essential to understanding the mind's causal powers.

Elisabeth is not satisfied. She immediately fires back:

> And I admit that it would be easier for me to concede matter and extension to the mind than it would be for me to concede the capacity to move a body and be moved by one to an immaterial thing.[21]

This is a remarkable statement; it may well be the first appearance of the causal argument for materialism (see chapter 4). For she is in effect saying that to allow for the possibility of mental causation, she would rather accept materialism concerning the mind ("it would be easier to concede matter and extension to the mind") than accept what she regards as an implausible dualist account offered by her mentor.

Why should anyone find Descartes's story so implausible? A couple of paragraphs back, it was pointed out that my mind's forming a "union" with my body amounts to the fact that my body serves as a necessary and omnipresent proximate cause and effect of changes in my mind and that my body is what makes it possible for my mind to have a causal influence on the outside world. Descartes, however, would reject this characterization of a mind-body union,

for the simple reason that it would beg the question as far as the possibility of mind-body causation is concerned. That is presumably why Descartes claimed that the notion of mind-body union is a "primitive"—one that is intelligible per se but is neither further explainable nor in need of an explanation. Should this answer have satisfied Elisabeth, or anyone else? A plausible case can be made for a negative answer. For when we ask what makes this body my body, not someone else's, a causal answer seems the most natural one and the only correct one. This is my body because it is the only body that I, or my desires and volitions, can directly move—that is, without moving or causally influencing anything else, whereas I can move other bodies, like this pen on my desk or the door to the hallway, only by moving my body first. Moreover, to cause any changes in my mind—or my mental states—you must first bring about appropriate changes in my body (presumably in my brain). What could be a more natural account of how my mind and my body form a "union"? But this explanation of mind-body union presupposes the possibility of mind-body causation, and it would be circular to turn around and say that an understanding of mind-body causation "depends" on the idea of mind-body union. Descartes's declaration that the idea of a union is a "primitive" and hence not in need of an explanation is unlikely to impress someone seeking an understanding of mental causation; it is liable to strike his critics simply as a dodge—a refusal to acknowledge a deep difficulty confronting his approach.

The "Pairing Problem": Another Causal Argument

We will develop another causal argument against Cartesian substance dualism. If this argument works, it will show not only that immaterial minds cannot causally interact with material things situated in space but also that they are not able to enter into causal relations with anything else, including other immaterial minds. Immaterial objects would be causally impotent and hence explanatorily useless; positing them would be philosophically unmotivated.

Here is the argument.[22] To set up an analogy and a point of reference, let us begin with an example of physical causation. A gun, call it *A*, is fired, and this causes the death of a person, *X*. Another gun, *B*, is fired at the same time (say, in *A*'s vicinity, but this is unimportant), and this results in the death of another person, *Y*. What makes it the case that the firing of *A* caused *X*'s death and the firing of *B* caused *Y*'s death, and not the other way around? That is, why did *A*'s firing not cause *Y*'s death and *B*'s firing not cause *X*'s death? What principle governs the "pairing" of the right cause with the right

effect? There must be a relation *R* that grounds and explains the cause-effect pairings, a relation that holds between *A*'s firing and *X*'s death and also between *B*'s firing and *Y*'s death, but not between *A*'s firing and *Y*'s death or between *B*'s firing and *X*'s death. What is this *R*, the "pairing relation," as we might call it? We are not necessarily supposing that there is a single such *R* for all cases of physical causation, only that some relation must ground the fact that a given cause is a cause of the particular effect that is caused by it.

Two ideas come to mind. First, there is the idea of a *causal chain*: There is a continuous causal chain connecting *A*'s firing with *X*'s death, as there is one connecting *B*'s firing with *Y*'s death, whereas no such chains exist between *A*'s firing and *Y*'s death or between *B*'s firing and *X*'s death. Indeed, with a high-speed video camera, we could trace the bullet's flight from each gun to its impact point on the target. The second idea is the thought that each gun when it fired was at a certain distance and in appropriate orientation in relation to the person it hit, but not to the other person. That is, *spatial relations* do the job of pairing causes with their effects.

A moment's reflection shows that the causal chain idea does not work as an independent solution to the problem. A causal chain, after all, is a series of events related as cause to effect, and interpolating more cause-effect pairs does not solve the pairing problem. For obviously it begs the question: What pairing relations ground these interpolated cause-effect pairs? It seems plausible that ultimately spatial relations—and more broadly, spatiotemporal relations—are the only way of generating pairing relations. Space appears to have nice causal properties; for example, as distance increases, causal influence diminishes, and it is often possible to set up barriers at intermediate positions to block or impede the propagation of causal influence. In any case, the following proposition seems highly plausible:

> (M) It is metaphysically possible for there to be two distinct physical objects, *a* and *b*, with the same intrinsic properties and hence the same causal potential or powers; one of these, say, *a*, causes a third object, *c*, to change in a certain way, but object *b* has no causal influence on *c*.

The fact that *a* but not *b* causes *c* to change must be grounded in some fact about *a*, *b*, and *c*. Since *a* and *b* have the same intrinsic properties, it must be their *relational properties* with respect to *c* that provide the desired explanation of their different causal roles. What relational properties or relations can do this job? It is plausible to think that when *a*, *b*, and *c* are physical objects, it is the spatial relation between *a* and *c* and that between *b* and *c* that are responsible

for the causal difference between *a* and *b* vis-à-vis *c*. (The object *a* was in the right spatial relation to *c*; *b* was "too far away" to exert any influence on it.) At least, there seems no other obvious candidate that comes to mind. Later we give an explanation of what it is about spatial relations that enables them to do the job.

Consider the possibility of immaterial souls, outside physical space, causally interacting with material objects in space. The following companion principle to (M) seems equally plausible, and if an interactionist substance dualist wishes to reject it, she should give a principled explanation why.

> (M*) It is metaphysically possible for there to be two souls, A and B, with the same intrinsic properties[23] such that they both act in a certain way at the same time and as a result a material object, C, undergoes a change. Moreover, it is the action of A, not that of B, that is the cause of the physical change in C.

What makes it the case that this is so? What pairing relation pairs the first soul, but not the second soul, with the material object? Since souls, as immaterial substances, are outside physical space and cannot bear spatial relations to anything, it is not possible to invoke spatial relations to ground the pairing. What possible relations could provide causal pairings across the two domains, one of spatially located material things and the other of immaterial minds outside space?

Consider a variation on the foregoing example: There are two physical objects, P_1 and P_2, with the same intrinsic properties, and an action of an immaterial soul causally affects one of them, say, P_1, but not P_2. How can we explain this? Since P_1 and P_2 have identical intrinsic properties, they must have the same causal capacity ("passive" causal powers as well as "active" causal powers), and it would seem that the only way to make them discernible in a causal context is their relations to other things. Doesn't that mean that any pairing relation that can do the job must be a spatial relation? If so, the pairing problem for this case is unsolvable since the soul is not in space and bears no spatial relation to anything. The soul cannot be any "nearer" to, or "more properly oriented" toward, one physical object than another. Nor could we say that there was a causal barrier "between" the soul and one of the physical objects but not the other, for what could "between" mean as applied to something in space and something outside it? It is a total mystery what nonspatial relations there could be that might help distinguish, from the point of view of an immaterial soul, between two intrinsically indiscernible physical objects.

Could there be causal interactions among immaterial substances? Ruling out mind-body causal interaction does not in itself rule out the possibility of a causally autonomous domain of immaterial minds in which minds are in causal commerce with other minds. Perhaps that is the picture of a purely spiritual afterlife envisioned in some religions and theologies. Is that a possibility? The pairing problem makes such an idea a dubious proposition. Again, any substance dualist who wants causation in the immaterial realm must allow the possibility of there being three mental substances, M_1, M_2, and M_3, such that M_1 and M_2 have the same intrinsic properties, and hence the same causal powers, and yet an action by M_1, but not the same action by M_2 at the same time, is causally responsible for a change in M_3. If such is a metaphysically possible situation, what pairing relation could connect M_1 with M_3 but not M_2 with M_3? If causation is to be possible within the mental domain, there must be an intelligible and motivated answer to this question. But what mental relations could serve this purpose? It is difficult to think of any.

Consider what space does for physical causation. In the kind of picture envisaged, where a physical thing or event causally acts on only one of the two objects with identical intrinsic properties, what distinguishes these two objects has to be their spatial locations with respect to the cause. Space provides a "principle of individuation" for material objects. Pure qualities and causal powers do not. And what enables space to serve this role is the fact that physical objects occupying exactly the same location in space at the same time are one and the same object.[24] This is in effect the venerable principle of "impenetrability of matter," which can usefully be understood as a sort of "exclusion" principle for space: Material things compete for, and exclude one another from, spatial locations. From this it follows that if physical objects *a* and *b* bear the same spatial relations to a third object *c*, *a* and *b* are one and the same object. This principle is what enables space to individuate material things with identical intrinsic properties. The same goes for causation in the mental domain. What is needed to solve the pairing problem for immaterial minds is a kind of mental coordinate system, a "mental space," in which these minds are each given a unique "location" at a time. Further, a principle of "impenetrability of minds" must hold in this mental coordinate system; that is, minds that occupy the same "location" in this space must be one and the same. It seems fair to say that we do not have any idea how a mental space of this kind could be constructed. Moreover, even if we could develop such a space for immaterial minds, that still would fall short of a complete solution to the pairing problem; to solve it for causal relations across the mental and physical domains, we need to somehow coordinate or fuse the two spaces, the

mental and the physical, to yield unitary pairing relations across the domains. It is not clear that we have any idea where to begin.

If there are Cartesian minds, therefore, they are threatened with total causal isolation—from each other as well as from the material world. The considerations presented do not show that causal relations cannot hold within a single mental substance (even Leibniz, famous for disallowing causation between monads, allowed it within a single monad). However, what has been shown seems to raise serious challenges for substance dualism. If this is right, we have a causal argument for a physicalist ontology. Causality requires a spacelike structure, and as far as we know, the physical domain is the only domain with a structure of that kind.

Immaterial Minds in Space?

All these difficulties with the pairing problem arise because of the radically nonspatial nature of minds in traditional substance dualism. According to Descartes, not only do minds lack spatial extension but also they are not in space at all. So why not bring minds into space, enabling them to have spatial locations and thereby solve the pairing problem? Most popular notions of minds as immaterial spirits do not seem to conceive them as wholly nonspatial. For example, when a person dies, her soul is thought to "rise" from the body, or otherwise "leave" it, implying that before the death the soul was inside the body and that the soul is capable of moving in space and changing its locations. Sometimes the departed souls of our loved ones are thought to be able to make their presence known to us in various ways, including in a visible form (think about Hamlet's ghostly father). It is probably impossible to make coherent sense of these popular ideas, but is there anything in principle wrong with locating immaterial minds in physical space and thereby making it possible for them to participate in the causal transactions of the world?

As we will see, the proposal to bring immaterial minds into space is fraught with complications and difficulties and probably not worth considering as an option. First there is the question of just where in space to put them. Is there a principled and motivated way of assigning a location to each soul? We might suggest that I locate my soul in my body, you locate your soul in your body, and so on. That may sound like a natural and reasonable suggestion, but it faces a number of difficulties. First, what about disembodied souls, souls that are not "united" with a body? Since souls are supposed to be substances in their own right, such souls are metaphysically possible. Second, if your soul is located in your body, exactly where in your body is it located? In

the brain, we might reply. But exactly where in the brain? It could not be spread all over the brain because minds are not supposed to be extended in space. If it has a location, the location has to be a geometric point. Is it coherent to think that there is a geometric point somewhere in your brain at which your mind is located? Descartes called the pineal gland the "seat of the soul," presumably because the pineal gland is where mind-body causal interaction was supposed to take place, although of course his official doctrine was that the soul is not in space at all.

Following Descartes's strategy here, however, does not seem to make much sense. For one thing, there is no evidence that there is any single place in the brain—a dimensionless point at that—at which mind-body interaction takes place. As far as we know, various mental states and activities are distributed over the entire brain and nervous system, and it does not make scientific sense to think, as Descartes did in regard to the pineal gland, that there is a single identifiable organ responsible for all mind-body causal interaction. Second, how could an entity occupying a single geometric point cause all the physical changes in the brain that are involved in mind-body causation? By what mechanism could this happen? How is energy transmitted from this geometric point to the neural fibers making up the brain? And there is this further question: What keeps the soul at that particular location? When I stand up from my chair in the study and go downstairs to the living room, somehow my soul tags along and moves exactly on the same trajectory as my body. When I board an airplane and the airplane accelerates on the runway and takes off, somehow my pointlike immaterial mind manages to gain speed exactly at the same rate and begins to cruise at the speed of 560 miles an hour! It seems that the soul is somehow firmly glued to some part of my brain and moves as my brain moves, and when I die it miraculously unglues itself from my body and migrates to a better (or perhaps worse) place in the afterlife. Does any of this make sense? Descartes was wise, we must conclude, to keep immaterial minds wholly outside physical space.

In any case, giving locations to immaterial minds will not in itself solve the pairing problem. As we saw, spatial locations of physical objects help solve the pairing problem in virtue of the principle that physical objects can be individuated in terms of their locations. As was noted, this is the principle of impenetrability of matter: Distinct objects exclude one another from spatial regions. That is how the causal roles of two intrinsically indiscernible physical objects could be differentiated. For the spatial locations of immaterial minds to help, therefore, we need a similar principle of spatial exclusion for immaterial minds—or the principle of impenetrability of mental substance—to the effect

that distinct minds cannot occupy exactly the same point in space. What reason is there to think such a principle holds? Why cannot a single point be occupied by all the souls that exist, like the thousand angels dancing on the head of a pin? Such a principle is needed if we are to make sense of causation for spatially located pointlike souls. But this does not mean that the principle is available; we must be able to produce independently plausible evidence or give a credible argument to show that the principle holds.

When we see all the difficulties and puzzles to which the idea of an immaterial mind, or soul, appears to lead, it is understandable why Descartes declared the notion of mind-body union to be primitive and not further explainable in terms of more fundamental ideas. Even a contemporary writer has invoked God and theology to make sense of how a particular mind (say, your mind) gets united to a particular body (your brain).[25] The reader is urged to think about whether such an appeal to theology gives us real help with the problems the dualist faces.

Substance Dualism and Property Dualism

It has seemed to most contemporary philosophers that the concept of mind as a mental substance is fraught with too many difficulties and puzzles without compensating explanatory gains. In addition, the idea of an immaterial and immortal soul usually carries with it various, often conflicting, religious and theological associations and aspirations that many of us would rather avoid in philosophical contexts. For example, the traditional conception of the soul involves a sharp and unbridgeable gap between humans and the rest of animal life. Even if our own mentality could be explained as consisting in the possession of a soul, what might explain the mentality of nonhuman animals? It is not surprising that substance dualism has not been a prominent alternative in contemporary philosophy of mind. But there is no call to exclude it a priori, without serious discussion; some highly reputable and respected philosophers continue to defend it as a realistic—perhaps the only—option (see "For Further Reading").

To reject the substantival view of mentality is not to deny that each of us "has a mind"; it is only that we should not think of "having a mind" literally—that is, as there being some object or substance called a "mind" that we literally possess. As discussed earlier (in chapter 1), having a mind is not like—at least, it need not be like—having brown eyes or a good throwing arm. To have

brown eyes, there must be brown eyes that you have. To "be out of your mind" or to "keep something in mind," you do not have to *have* some object—namely, a mind—which you are out of, or in which you keep something. If you have set aside substance dualism, at least for now, you can take having a mind simply as having a certain special set of *properties, capacities,* and *characteristics*, something that humans and some higher animals possess but sticks and stones do not. To say that something "has a mind" is to classify it as a certain sort of thing—as a thing with capacities for certain characteristic sorts of behavior and functions, such as sensation, perception, memory, learning, consciousness, and goal-directed action. For this reason, it is less misleading to speak of "having mentality" than "having a mind." (As you will recall, this is what the last dualist argument we considered above, "Leibniz's mill," challenges; the point of the argument is precisely that no material system can have mentality.)

In any case, substance dualism has played a small role in contemporary philosophy of mind. Philosophical attention has focused instead on mental activities and functions—or mental events, states, and processes—and the mind-body problem has turned into the problem of understanding how these mental events, states, and processes are related to physical and biological events, states, and processes, or how our mental or psychological capacities and functions are related to the nature of our physical structure and capacities. In regard to this question, there are two principal positions: *property dualism* and *reductive physicalism* (also called *type physicalism*). Dualism is no longer a dualism of two sorts of substances; it is now a dualism of two sorts of properties, mental and physical. "Property" is used here in a broad sense: Mental properties comprise mental functions, capacities, events, states, and the like, and similarly for physical properties. It is a catchall term referring to events, activities, states, and the rest. So property dualism is the view that mental properties are diverse from and irreducible to physical properties. In contrast, reductive physicalism defends the position that mental properties are reducible to, and therefore can be identified with, physical properties. As we will see, there are various forms of both property dualism and reductionist physicalism. However, they all share one thing in common: the rejection of immaterial minds. Contemporary property dualism and reductive physicalism acknowledge only objects of one kind in the world—bits of matter and increasingly complex structures aggregated out of bits of matter. (This anti-Cartesian position is called substance physicalism.) Some of these physical systems exhibit complex behaviors and activities, like perceiving, sensing, reasoning, and consciousness. But these are only properties of material structures. The main point of dispute concerns the

nature of the relationship between these mental features and activities on one hand and the structures' physical characteristics on the other. This is the central question for the remainder of this book.

For Further Reading

The primary source of Descartes's dualism is his *Meditations on First Philosophy*, first published in 1641. See especially Meditations II and VI. There are numerous English editions; a good version (including *Objections and Replies*) can be found in *The Philosophical Writings of Descartes*, vol. 2, translated and edited by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Helpful historical and interpretive literature on Descartes's philosophy of mind includes: Daniel Garber, *Descartes Embodied* (especially chapter 8, "Understanding Causal Interaction: What Descartes Should Have Told Elisabeth"); Marleen Rozemond, *Descartes's Dualism*, chapter 1; and Lilli Alanen, *Descartes's Concept of Mind*, chapter 2.

On the pairing problem, see Kim, *Physicalism, or Something Near Enough*, chapter 3. For dualist responses, John Foster, "A Defense of Dualism"; Andrew Baily, Joshua Rasmussen, and Luke Van Horn, "No Pairing Problem."

For some contemporary defenses of substance dualism, see John Foster, *The Immaterial Self*; W. D. Hart, *The Engines of the Soul*; William Hasker, *The Emergent Self*; E. J. Lowe, "Non-Cartesian Substance Dualism and the Problem of Mental Causation," and "Dualism"; Alvin Plantinga, "Against Materialism"; Dean Zimmerman, "Material People"; and Richard Swinburne, *The Evolution of the Soul*.

Also recommended are Noa Latham, "Substance Physicalism," and Tim Crane, "Mental Substances."

Notes

1. See, for example, John Foster, *The Case for Idealism*.

2. Descartes writes: "Substance: this term applies to every thing in which whatever we perceive immediately resides, as in a subject. . . . By 'whatever we perceive' is meant any property, quality or attribute of which we have a real idea." See "Author's Replies to the Second Set of Objections," p. 114.

3. René Descartes, "Author's Replies to the Fourth Set of Objections," p. 159.

4. Many philosophers in Descartes's time, including Descartes himself, held that, strictly speaking, God is the only being capable of independent ex-

istence and therefore that the only true substance is God, all others being "secondary" or "derivative" substances.

5. René Descartes, *Meditations on First Philosophy*, Meditation II, p. 19.

6. One might say that this is only a case of physical possibility and necessity, not possibility and necessity *tout court*. A more standard example would be the proposition that water = H_2O. It is widely accepted that this is a necessary truth (though a posteriori) but that its falsehood is conceivable.

7. See some of the essays in *Conceivability and Possibility*, edited by Tamar Szabo Gendler and John Hawthorne. Gendler and Hawthorne's introduction is a good starting point.

8. This approach, called "animalism," has recently been receiving much attention. See, for example, Eric T. Olson, *The Human Animal: Personal Identity Without Psychology*.

9. Strictly, (NI) holds only when X and Y are "rigid designators." A name is said to be "rigid" just in case it names the same thing in every possible world in which it exists. In this sense, "Cicero" and "Tully," along with most proper names, are rigid. For details, see Saul Kripke, *Naming and Necessity*.

10. John Locke, *An Essay Concerning Human Understanding*, Book II, chapter 27, secs. 15, 16.

11. As noted by Marleen Rozemond in her *Descartes's Dualism*, p. 3.

12. Gottfried Leibniz, *Monadology*, 17.

13. Alvin Plantinga, "Against Materialism," p. 13.

14. John Foster, "A Brief Defense of the Cartesian View," pp. 25–26. "The kinds of mentality which the dualist postulates" refers to mentality conceived as irreducible to physical processes. Foster of course believes that mentality cannot be physically reduced; the point is that if mental states are reduced to, say, neural states of an organism, there would be no special problem about how material things can have mentality.

15. Functionalism (chapters 5, 6) can be seen as providing a story that explains how physical systems can have beliefs, desires, emotions, and so on. As we will see, functionalism construes mental states as "functional states," that is, states defined in terms of the causal work they perform. Such states are "realized" by states in physical systems and it is claimed that these physical realizers do the causal work required for intentional states. Thus, a physical system has a certain belief when one of its physical states realizes the belief. See also chapter 10 on David Chalmers on the "hard" and "easy" problems of consciousness. Dualists like Plantinga will reject the claim that mental states are functional states.

16. René Descartes, *Meditations on First Philosophy*, Meditation VI, pp. 59–60.

17. René Descartes, *The Passions of the Soul*, I, 41, p. 343.

18. Daniel Garber, "Understanding Interaction: What Descartes Should Have Told Elisabeth," p. 172. This and other quotations from the correspondence between Elisabeth and Descartes are taken from this chapter of Garber's book, *Descartes Embodied*.

19. René Descartes, *Meditations on First Philosophy*, Meditation II, p. 17.

20. Descartes to Princess Elisabeth, May 21, 1643, in Garber, *Descartes Embodied*, p. 173.

21. Princess Elisabeth to Descartes, June 1643, in Garber, *Descartes Embodied*, p. 172.

22. For a fuller presentation of this argument, see Kim, *Physicalism, or Something Near Enough*, chapter 2. For some dualist responses, see the "For Further Reading" section.

23. If you are inclined to invoke the identity of intrinsic indiscernibles for souls to dissipate the issue, the next situation we consider involves only one soul and this remedy does not apply. Moreover, the pairing problem can be generated without assuming that there can be distinct intrinsic indiscernibles. This assumption, however, helps to present the problem in a simple and compelling way.

24. There is the familiar problem of the statue and the lump of clay of which it is composed (the problem of coincident objects). Some claim that although these occupy the same region of space and coincide in many of their properties (for example, weight, shape, size), they are distinct objects because their persistence conditions are different (for example, if the clay is molded into a cube, the clay, but not the statue, continues to exist). We must set this problem aside, but it does not affect our argument. Note that the statue and the lump of clay share the same causal powers and suffer the same causal fate (except perhaps coming into being and going out of existence).

25. John Foster, "A Brief Defense of the Cartesian View."

CHAPTER 3

Mind and Behavior

Behaviorism

Behaviorism arose early in the twentieth century as a doctrine on the nature and methodology of psychology, in reaction to what some psychologists took to be the subjective and unscientific character of introspectionist psychology. In his classic *Principles of Psychology*, published in 1890, William James, who had a major role in establishing psychology as a scientific field, begins with an unambiguous statement of the scope of psychology:

> Psychology is the Science of Mental Life, both of its phenomena and of their conditions. The phenomena are such things as we call feelings, desires, cognitions, reasonings, decisions, and the like.[1]

For James, then, psychology was the scientific study of mental phenomena, with the study of conscious mental processes as its core task. As for the method of investigation of these processes, James writes: "Introspective observation is what we have to rely on first and foremost and always."[2]

Compare this with the declaration in 1913 by J. B. Watson, who is considered the founder of the behaviorist movement: "Psychology . . . is a purely objective experimental branch of natural science. Its theoretical goal is the prediction and control of behavior."[3]

This view of psychology as an experimental study of publicly observable human and animal behavior, not of inner mental life observed through private introspection, dominated scientific psychology and associated fields until the 1960s and made "behavioral science" a preferred name for psychology in universities and research centers around the world, especially in North America.

The rise of behaviorism and the influential position it attained was no fluke. Even James saw the importance of behavior to mentality; in *The Principles of Psychology*, he also writes:

> *The pursuance of future ends and the choice of means for their attainment are thus the mark and criterion of the presence of mentality* in a phenomenon. We all use this test to discriminate between an intelligent and a mechanical performance. We impute no mentality to sticks and stones, because they never seem to move for *the sake of* anything.[4]

It is agreed on all sides that behavior is intimately related to mentality. Obviously, what we do is inseparably connected with what we think and want, how we feel, and what we intend to accomplish. Our behavior is a natural expression of our beliefs and desires, feelings and emotions, and goals and aspirations. But what precisely is the relationship? Does behavior merely serve, as James seems to be suggesting, as an *indication*, or a *sign*, that a mind is present? And if behavior is a sign of mentality, what makes it so? If something serves as a sign of something else, there must be an underlying relationship that explains why the first can serve as a sign of the second. Fall in the barometric pressure is a sign of an oncoming rain; that is based on observed regular sequences. Is behavior related to minds in a similar way? Not likely: You can wait and see if rain comes; you presumably can't look inside another mind to see if it's really there!

Or is the relationship between behavior and mentality a more intimate one? Philosophical behaviorism takes behavior as *constitutive* of mentality: Having a mind just *is* a matter of exhibiting, or having a *propensity* or *capacity* to exhibit, appropriate patterns of behavior. Although behaviorism, in both its scientific and philosophical forms, has lost the sweeping influence it once enjoyed, it is a doctrine that we need to understand in some depth and detail, since not only does it form the historical backdrop of much of the subsequent thinking about the mind, but its influence lingers on and can be discerned in some important current philosophical positions. In addition, a proper appreciation of its motivation and arguments will help us gain a better understanding of the relationship between behavior and mentality. As we will see, it cannot be denied that behavior has something crucial to do with minds, although this relationship may not have been correctly conceived by behaviorism. Further, reflections on the issues that motivated behaviorism can help us gain an informed perspective on the nature and status of psychology and cognitive science.

The Cartesian Theater and the "Beetle in the Box"

On the traditional conception of mind deriving from Descartes, the mind is a private inner stage, aptly called the Cartesian theater by some philosophers,[5] on which mental actions take place. It is the arena in which our thoughts, bodily sensations, perceptual sensings, volitions, emotions, and all the rest make their appearances, play out their assigned roles, and then fade away. All this for an audience of one: One and only one person has a view of the stage, and no one else is permitted a look. Moreover, that single person, who "owns" the theater, has a full and authoritative view of what goes in the theater: Nothing that appears on the stage escapes her notice. She is in total cognitive charge of her theater. In contrast, the outsiders must depend on what she says and does to guess what might be happening in the theater; no direct viewing is allowed.

I know, directly and authoritatively, that I am having a pain in my bleeding finger. You can see the bleeding finger, and hear my words "Oh damn! This hurts!" and come to believe that I must be experiencing a bad pain. Your knowledge of my pain is based on observation and evidence, though probably not explicit inference, whereas my knowledge of it is direct and immediate. You see your roommate leaving the apartment with her raincoat on and carrying an umbrella, and you reason that she thinks it is going to rain. But she knows what she thinks without having to observe what she is doing with her raincoat; she knows it directly. Or so it seems. Evidently, all this points to an asymmetry between the first person and the third person where knowledge of mental states is concerned: Our knowledge of our own current mental states is *direct*, in that it is not mediated by evidence or inference, and *authoritative*, or *privileged*, in that in normal circumstances, it is immune to the third person's challenge, "How do you know?" This question is a demand for evidence for your knowledge claim. Since your knowledge is not based on evidence, or inference from evidence, there is nothing for you to say, except perhaps "I just know."

Early in the twentieth century, however, some philosophers and psychologists began to question this traditional conception of mentality; they thought that it led to unacceptable consequences, consequences that seemingly contradict our ordinary assumptions and practices involving knowledge of other minds and our use of language to talk about mental states, both ours and others'.

The difficulty is *not* that such knowledge, based as it is only on "outer" signs, is liable to error and cannot attain the kind of certainty with which we supposedly know our own minds. The problem, as some saw it, goes deeper:

It makes knowledge of other minds not possible at all! Take a standard case of inductive inference—inference based on premises that are less than logically conclusive—such as this: You find your roommate listening to the weather report on the radio, which is predicting heavy showers later in the day, and say to yourself, "She is going to be looking for her umbrella!" This inference is liable to error: Perhaps she misunderstood the weather report or wasn't paying attention, or she rather enjoys getting wet. Now compare this with our inference of a person's pain from her "pain behavior." There is this difference: In the former case, you can check by further observation whether your inference was correct (you can wait and see whether she looks for her umbrella), but with the latter, further observation yields only more observation of her behavior, never an observation of her pain! Only she can experience her pains; all you can do is to see what she does and says. And what she *says* is only behavior of another kind. (Maybe she is very stoic and reserved about little pains and aches.) One hallmark of induction is that inductive predictions can be confirmed or disconfirmed—you just wait and see whether the predicted outcome occurs. For this reason, inductive procedures are said to be self-correcting; predictive successes, or lack thereof, are their essential constraint. In contrast, predictions of inner mental events on behavioral evidence cannot be verified one way or the other, and not subject to correction. As a result, there is no predictive constraint on them. This makes it dubious whether these are legitimate inferences from behavior to inner mental states at all.

The point is driven home by Ludwig Wittgenstein's parable of "the beetle in the box." Wittgenstein writes:

> Suppose everyone had a box with something in it; we call it a "beetle." No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at *his* beetle. Here it would be quite possible for everyone to have something different in his box.[6]

As it happens, you have a beetle in your box, and everyone else says that they too have a beetle in their box. But what can you know from their utterances, "I have a beetle in my box"? How would you know what they mean by the word "beetle"?

The apparent answer is that there is no way for you to know what others mean by "beetle," or to confirm whether they have in their boxes what you have in yours: For all you know, some may have a butterfly, some may have a little rock, and perhaps others have nothing at all in their boxes. Nor can others know what you mean when they hear you say, "I have a beetle in my box."

As Wittgenstein says, the thing in the box "cancels out whatever it is." It is difficult to see how the word "beetle" can have a common meaning that can be shared by speakers, or how the word "beetle" could have a role in the exchange of information.

A deeper lesson of Wittgenstein's beetle, therefore, is that it is mysterious how, on the Cartesian conception of the mind, we could ever fix the meaning of the word "pain" and use utterances like "I have a pain in my knee" to impart information to other speakers. For the pain case seems exactly analogous to the beetle in the box: Suppose you and your friends take a fall while running on the track and all of you bruise your knees. Everyone cries out "My knee hurts!" On the Cartesian picture, something is going on in each person's mind, but each can observe only what's going on in her mind, not what's going on in anyone else's. Is there any reason to think that there is something common, some identical sensory experience, going on in everyone's mind, in each Cartesian theater? Pain in the mind seems just as elusive as the beetle in the box. You are experiencing pain; another person could be feeling an itch in the knee; still others could have a tickle; some may be having a sensation unlike anything you have ever experienced; and some may not be having any sensation at all. As Wittgenstein would have said, the thing in each mind cancels out whatever it is.

Evidently, however, we use utterances like "My knee hurts" to communicate information to other people, and expressions like "pain" and "the thought that it's going to rain" have intersubjective meanings, meanings that can be shared by different speakers. Your pain gets worse and you decide to go to a clinic. Gently tapping your kneecap with her fingers, your physician asks, "Does it hurt?" You reply, "Yes, it does, Doctor." This is a familiar kind of exchange in a medical office, and it can be important to diagnosis and treatment. But the exchange makes no sense unless the words "the knee hurts" on your doctor's mouth mean the same as "the knee hurts" on your mouth; unless the expression has a shared meaning for you and your doctor, your reply could not count as an answer to your doctor's question. You and your doctor would be talking past each other. Our psychological language, the language in which we talk about sensations, likes and dislikes, hopes and regrets, thoughts, emotions, and the rest, is an essential vehicle of social interchange and interaction; without a language in which we communicate with each other about such matters, social life as we know it is scarcely imaginable. For this to be possible, the expressions of this language must have by and large stable and invariant meanings from speaker to speaker. What we have seen is that the privacy of the Cartesian minds may well infect psychological language, making it essentially private as

well. The problem is that a private language fails as a genuine language, because the defining function of language is to serve as an instrument of interpersonal communication. All this seems to discredit the Cartesian picture of the mind as an inner theater for an audience of one.

Behaviorism is a response to these seemingly unacceptable consequences of the Cartesian conception of the mind. It rejects the traditional picture of how our mental expressions acquire their meanings by referring to private inner episodes, and attempts to ground their meanings in publicly accessible and verifiable facts and conditions about people. According to the behaviorist approach, the meanings of mental expressions, such as "pain" and "thought," are to be explained by reference to facts about observable behavior—how people who have pain or thoughts act and behave. But what is meant by "behavior"?

What Is Behavior?

As our first pass, we can take "behavior" to mean whatever people or organisms, or even mechanical systems, *do* that is *publicly observable*. "Doing" is to be distinguished from "having something done," though this distinction is not always clear. If you grasp my arm and pull it up, the rising of my arm is not something I do; it is not my behavior (but your pulling up my arm is behavior—your behavior). It is not something that a psychologist would be interested in investigating. But if I raise my arm—that is, if I cause it to rise—then it is something I do, and it counts as my behavior. It is not assumed here that the doing must in some sense be "intentional" or done for a purpose; it is only required that it is proximately caused by some occurrence internal to the behaving system. If a robot moves toward a table and picks up a book, its movements are part of its behavior, regardless of whether the robot "knows" or "intends" what it is doing. If a bullet punctures the robot's skin, that is not part of its behavior, not something it does; it is only something that happens to it.[7]

What are some examples of things that humans and other behaving organisms do? Let us consider the following four possible types:

i. *Physiological reactions and responses*: for example, perspiration, salivation, coughing, increase in the pulse rate, rising blood pressure.[8]
ii. *Bodily movements*: for example, walking, running, raising a hand, opening a door, throwing a baseball, a cat scratching at the door, a rat turning left in a T-maze.

iii. *Actions involving bodily motions*: for example, greeting a friend, writing an e-mail, going shopping, writing a check, attending a concert.
iv. *Actions not involving overt bodily motions*: for example, judging, reasoning, guessing, calculating, deciding, intending.

Behaviors falling under (iv), sometimes called "mental acts," evidently involve "inner" events that cannot be said to be publicly observable, and behaviorists do not consider them "behavior" in their sense. (This, however, does not necessarily rule out behavioral interpretations of these activities.) Those falling under (iii), although they involve bodily movements, also have clear and substantial psychological components. Consider the act of writing a check: Only if you have certain cognitive capacities, beliefs, desires, and an understanding of relevant social institutions can you write a check. You must have a desire to make a payment and the belief that writing a check is a means toward that end. You must also have some understanding of exchange of money for goods and services and the institution of banking. The main point is this: A person whose observable behavior is indistinguishable from yours when you are writing a check is not necessarily writing a check, and a person who is waving his hand just like you are waving yours may not be greeting a friend although you are (try to think how these things can happen). Something like this is true of other examples listed under (iii), and this means that none of these count as behavior for the behaviorist. Remember: Public observability is key to the behaviorist conception of behavior. This implies that if two behaviors are observationally indistinguishable, they must count as the "same" behavior.

So only those behaviors under (i) and (ii) on our list—what some behaviorists called "motions and noises"—meet the behaviorist requirements. In much behaviorist literature, there is an assumption that only physiological responses and bodily motions that are in a broad sense "overt" and "external" are to count as behavior. This could rule out events and processes occurring in the internal organs; thus, internal physiological states, including states of the brain, would not, on this view, count as behavior, although they are physical states and conditions that are intersubjectively accessible. The main point to remember, though, is that however the domain of behavior is circumscribed, behavior is taken to be bodily events and conditions that are publicly accessible to all competent observers. Behavior in this sense does not enjoy the kind of privileged access granted to the first person in the Cartesian picture. That is, *equal access for all* is of the essence of behavior as conceived by the behaviorist.

Logical Behaviorism: A Positivist Argument

Writing in 1935, Carl G. Hempel, a leading logical positivist, said, "We see clearly that the meaning of a psychological statement consists solely in the function of abbreviating the description of certain modes of physical response characteristic of the bodies of men and animals."[9]

This is what is called "logical behaviorism," because it is based on the supposed close logical connections between psychological expressions and expressions referring to behavior. It is also called "analytical behaviorism" or "philosophical behaviorism" (to be distinguished from scientific, or methodological behaviorism; see below). Fundamentally, it is a claim about the translatability of psychological sentences into sentences that ostensibly refer to no inner psychological occurrences but only to publicly observable aspects of the subject's behavior and physical conditions. More formally, the claim can be stated like this:

> *Logical Behaviorism I*. Any meaningful psychological statement, that is, a statement purportedly describing a mental phenomenon, can be *translated*, without loss of content, into a cluster of statements solely about behavioral and physical phenomena.

And the claim can be formulated somewhat more broadly as a thesis about the behavioral definability of all meaningful psychological expressions:

> *Logical Behaviorism II*. Every meaningful psychological expression can be *defined* solely in terms of behavioral and physical expressions, that is, expressions referring to behavioral and physical phenomena.

Here "definition" is to be understood in the following fairly strict sense: If an expression E is defined as E^*, then E and E^* must be either synonymous or conceptually equivalent (that is, as a matter of meaning, there is no conceivable situation to which one of the expressions applies but the other does not).[10] Assuming translation to involve synonymy or at least conceptual equivalence, we can see that logical behaviorism (II) entails logical behaviorism (I).

Why should anyone accept logical behaviorism? The following argument extracted from Hempel represents one important line of thinking that led to the behaviorist position:

1. The meaning of a sentence is given by the conditions that must be verified to obtain if the sentence is true (we may call these "verification conditions").
2. If a sentence has a meaning that can be shared by different speakers, its verification conditions must be accessible to each speaker—that is, they must be publicly observable.
3. Only behavioral and physical phenomena (including physiological occurrences) are publicly observable.
4. Therefore, the sharable meaning of any psychological sentence must be specifiable by statements of publicly observable verification conditions, that is, statements describing behavioral and physical conditions that must hold if the psychological statement is true.

Premise (1) is called "the verifiability criterion of meaning," a central doctrine of the philosophical movement of the early twentieth century known as logical positivism. The idea that meanings are verification conditions is no longer widely accepted, though it is by no means dead. However, we can see and appreciate the motivation to go for something like the intersubjective verifiability requirement in the following way. We want our psychological statements to have public, sharable meanings and to serve as vehicles of interpersonal communication. Suppose someone asserts a sentence S. For me to understand what S means, I must know what state of affairs is represented by S (for example, whether S represents snow's being white or the sky's being blue). But for me to know what state of affairs this is, it must be one that is accessible to me; it must be the kind of thing that I could in principle determine to obtain or not to obtain. It follows that if the meaning of S—namely, the state of affairs that S represents—is to be intersubjectively sharable, it must be specified by conditions that are intersubjectively accessible. Therefore, if psychological statements and expressions are to be part of public language suitable for intersubjective communication, their meanings must be governed by publicly accessible criteria, and only behavioral and physical conditions qualify as such criteria. And if anyone insists that there are inner subjective criteria for psychological expressions as well, we should reply, the behaviorist would argue, that even if such existed, they (like Wittgenstein's beetles) could not be part of the meanings that can be understood and shared by different persons. Summarizing all this, we could say: Insofar as psychological expressions have interpersonal meanings, they must be definable in terms of behavioral and physical expressions.

A Behavioral Translation of "Paul Has a Toothache"

As an example of behavioral and physical translation of psychological statements, let us see how Hempel proposes to translate "Paul has a toothache" in behavioral terms. His translation consists of the following five clauses:[11]

a. Paul weeps and makes gestures of such and such kinds.
b. At the question "What is the matter?" Paul utters the words, "I have a toothache."
c. Closer examination reveals a decayed tooth with exposed pulp.
d. Paul's blood pressure, digestive processes, the speed of his reactions, show such and such changes.
e. Such and such processes occur in Paul's central nervous system.

Hempel suggests that we regard this list as open-ended; there may be many other such "test sentences" that would help to verify the statement that Paul is having a toothache. But how plausible is the claim that these sentences together constitute a behavioral-physical translation of "Paul has a toothache"?

It is clear that as long as translation is required to preserve "meaning" in the ordinary sense, we must disqualify (d) and (e): It is not a condition on the mastery of the meaning of "toothache" that we know anything about blood pressure, reaction times, and conditions of the nervous system. Even (c) is questionable: Why can't someone experience toothache (that is, have a "toothachy" pain) without having a decayed tooth or in fact any tooth at all? (Think about "phantom pains" in an amputated limb.) (If "toothache" means "pain caused by an abnormal physical condition of a tooth," then "toothache" is no longer a purely psychological expression.) This leaves us with (a) and (b).

Consider (b): It associates *verbal behavior* with toothache. Unquestionably, verbal reports play an important role in our finding out what other people are thinking and feeling, and we might think that verbal reports, and verbal behavior in general, are observable behavior that we can depend on for knowledge of other minds. But there is a problem: Verbal behavior is not pure physical behavior, behavior narrowly so called. In fact, it can be seen that verbal behavior, such as responding to a question with an utterance like "I have a toothache," presupposes much that is robustly psychological; it is a behavior of kind (iv) distinguished earlier. For Paul's response to be relevant here, he must *understand* the question "What is the matter?" and *intend to express the belief* that he has a toothache, by uttering the sentence "I have a toothache."

Understanding a language and using it for interpersonal communication is a sophisticated, highly complex cognitive ability, not something we can subsume under "motions and noises." Moreover, given that Paul is having a toothache, he responds in the way indicated in (b) *only if he wants to tell the truth*. But "want" is a psychological term, and building this clause into (b) would again compromise its behavioral-physical character. We must conclude that (b) is not an eligible behavioral-physical "test sentence." We return to some of these issues in the next section.

Difficulties with Behavioral Definitions

Let us consider beliefs: How might we define "S believes that there are no native leopards in North America" in terms of S's behavior? Pains are associated with a rough but distinctive range of behavior patterns, such as winces, groans, screams, characteristic ways in which we favor the affected bodily parts, and so on, which we may collectively call "pain behavior" (recall Hempel's condition [a]). However, it is much more difficult to associate higher cognitive states with specific patterns of behavior. Is there even a loosely definable range of bodily behavior that is characteristically and typically exhibited by all people who believe that there are no native leopards in North America, or that free press is essential to democracy? Surely the idea of looking for bodily behaviors correlated with these beliefs makes little sense.

This is why it is tempting, perhaps necessary, to resort to the idea of *verbal behavior*—the disposition to produce appropriate verbal responses when prompted in certain ways. A person who believes that there are no native leopards in North America has a certain linguistic disposition—for example, he would tend to utter the sentence "There are no native leopards in North America," or its synonymous variants, under certain conditions. This leads to the following schematic definition:

> S believes that $p =_{\text{def}}$ If S is asked, "Is it true that *p*?" S will answer, "Yes, it is true that *p*."

The right-hand side of this formula (the "definiens") states a *dispositional* property (*disposition* for short) of S: S has a disposition, or propensity, to produce behavior of an appropriate sort under specified conditions. It is in this sense that properties like being soluble in water or being magnetic are called dispositions: Water-soluble things dissolve when immersed in water, and magnetic objects attract iron filings that are placed nearby. To be soluble at time *t*, it

need not be dissolving at *t*, or ever. To have the belief that *p* at time *t*, you only need to be disposed, at *t*, to respond appropriately if prompted in certain ways; you need not actually produce any of the specified responses at *t*.

There is no question that something like the above definition plays a role in finding out what other people believe. And it should be possible to formulate similar definitions for other propositional attitudes, like desiring and hoping. The importance of verbal behavior in the ascription of beliefs can be seen when we reflect on the fact that we are willing to ascribe to nonverbal animals only crude and rudimentary beliefs. We routinely attribute to a dog beliefs like "The food bowl is empty" and "There is a cat sitting on the fence," but not beliefs like "Either the food bowl is empty or there is no cat sitting on the fence" and "If no cat is sitting on the fence, either it's raining or his master has called him in." It is difficult to think of nonverbal behavior on the basis of which we can attribute to anyone, let alone cats, beliefs with logically complex contents, say, beliefs expressed by "Every cat can be fooled some of the time, but no cat can be fooled all of the time," or "Since tomorrow is Monday, my master will head for work in Manhattan as usual, unless his cold gets worse and he decides to call in sick," and the like. It is arguable that in order to have beliefs or entertain thoughts like these, you must be a language user with a capacity to generate and understand sentences with complex structure.

Confining our attention to language speakers, then, let us see how well the proposed definition of belief works as a behaviorist definition. Difficulties immediately come to mind. First, as we saw with Hempel's "toothache" example, the definition presupposes that the person in question *understands* the question "Is it the case that *p*?"—and understands it *as a request for* an answer of a certain kind. (The definition as stated presupposes that the subject understands English, but this feature of the definition can be eliminated by modifying the antecedent, thus: "S is asked a question in a language S understands that is synonymous with the English sentence 'Is it the case that *p*?'") But understanding is a psychological concept, and if this is so, the proposed definition cannot be considered behavioristically acceptable (unless we have a prior behavioral definition of "understanding" a language). The same point applies to the consequent of the definition: In uttering the words "Yes, it is the case that *p*," S must *understand what these words mean* and *intend them to be understood by her hearer to have that meaning*. It is clear that speech acts like saying something and uttering words with an intention to communicate carry substantial psychological presuppositions about the subject. If they are to count as "behavior," it would seem that they must be classified as type (iii) or (iv) behavior, not as motions and noises.

A second difficulty (this too was noted in connection with Hempel's example): When S is asked the question "Is it the case that *p*?" S responds in the desired way only if S *wants* to tell the truth. Thus, the condition "if S wants to tell the truth" must be added to the antecedent of the definition, but this again threatens its behavioral character. The belief that *p* leads to an utterance of a sentence expressing *p* only if we combine the belief with a certain desire, the desire to tell the truth. The point can be generalized: Often behavior or action issues from a complex of mental states, not from a single, isolated mental state. As a rule, beliefs alone do not produce any specific behavior unless they are combined with appropriate desires.[12] Nor will desires: If you want to eat a ham sandwich, this will lead to your ham-sandwich-eating behavior only if you believe that what you are handed is a ham sandwich; if you believe that it is a beef-tongue sandwich, you may very well pass it up. If this is so, it seems not possible to define belief in behavioral terms without building desire into the definition, and if we try to define desire behaviorally, we find that that is not possible unless we build belief into *its* definition.[13] This would indeed be a very small definitional circle.

The complexity of the relationship between mental states and behavior can be appreciated in a more general setting. Consider the following schema relating desire, belief, and action:

> *Desire-Belief-Action Principle (DBA)*. If a person desires that *p* and believes that doing A is an optimal way to secure that *p*, she will do A.

There are various ways of sharpening this principle: For example, it is probably more accurate to say, "She will try to do A" or "She will be disposed to do A," rather than "She will do A." In any event, some such principle as DBA underlies our "practical reasoning"—the means-ends reasoning that issues in action. It is by appeal to such a principle that we "rationalize" actions—that is, give reasons that explain why people do what they do. DBA is also useful as a predictive tool: When we know that a person has a certain desire and that she takes a certain action as an effective way of securing what she desires, we can reasonably predict that she will do, or try to do, the required action. Something like DBA is often thought to be fundamental to the very concept of "rational action."

Consider now an instance of DBA:

1. If Mary desires that fresh air be let into the room and believes that opening the window is a good way to make that happen, she will open the window.

Is (1) true? If Mary does open the window, we could explain her behavior by appealing to her desire and belief as specified in (1). But it is clear that she may have the desire and belief but not open the window—not if, for example, she thinks that opening the window will also let in the horrible street noise that she abhors. So perhaps we could say:

2. If Mary desires fresh air to be let in and believes that opening the window is a good way to make that happen, but if she also believes that opening the window will let in the horrible street noise, she will not open the window.

But can we count on (2) to be true? Even given the three antecedents of (2), Mary will still open the window if she also believes that her ill mother very badly needs fresh air. It is clear that this process could go on indefinitely.

This suggests something interesting and very important about the relationship between mental states and behavior, which can be stated like this:

Defeasibility of Mental-Behavioral Entailments. If there is a plausible entailment of behavior B by mental states $M_1, \ldots, M_n$, there always is a further mental state M_{n+1} such that $M_1, \ldots, M_n, M_{n+1}$ together plausibly entail not-B.

If we assume not-B (that is, the failure to produce behavior B) to be behavior as well, the principle can be iteratively applied, without end, as we saw with Mary and the window opening: There exists some mental state M_{n+2} such that $M_1, \ldots, M_n, M_{n+1}, M_{n+2}$ together plausibly entail B. And so on without end.[14]

This shows that the relationship between mental states and behavior is highly complex: The moral is that mind-to-behavior connections are always *defeasible*—and defeasible by the occurrence of a *further mental state*, not merely by physical barriers and hindrances (as when Mary cannot open the window because her arms are paralyzed or the window is nailed shut). This makes the prospect of producing for each mental expression a purely behavioral-physical definition extremely remote. But we should not lose sight of the important fact that the defeasibility thesis does state an important and interesting connection between mental phenomena and behavior. The thesis does not say that there are no mental-behavioral entailments—it only says that such entailments are more complex than they might first appear, in that they always face potential mental defeaters.

Let us now turn to another issue. Suppose you want to greet someone. What behavior is entailed by this want? As we might say, greeting desires issue in greeting behavior. But what is greeting behavior? When you see Mary across the street and want to greet her, you might wave to her, cry out "Hi, Mary!" The entailment is defeasible since you would not greet her, even though you want to, if you also thought that by doing so you might cause her embarrassment. Be that as it may, saying that wanting to greet someone issues in a *greeting* does not say much about the *observable physical behavior*, because greeting is an action that includes a manifest psychological component (behavior of type [iii] distinguished earlier). Greeting Mary involves *noticing* and *recognizing* her, *believing* (or *hoping*) that she will *notice* your physical gesture and *recognize* it as expressing your *intention* to greet her, and so on. Greeting obviously will not count as behavior of kind (i) or (ii)—that is, a physiological response or bodily movement.

But does wanting to greet entail any bodily movements? If so, what bodily movements? There are innumerable ways of greeting: You can greet by waving your right hand, waving your left hand, or waving both; by saying "Hi!" or "How are you?" or "Hey, how're you doing, Mary?"; by saying these things in French or Chinese (Mary is from France, and you and Mary are taking a Chinese class); by rushing up to Mary and shaking her hand or giving her a hug; and countless other ways. In fact, any physical gesture will do as long as it is socially recognized as a way of greeting.[15]

And there is a flip side to this. As travel guidebooks routinely warn us, a gesture that is recognized as friendly and respectful in one culture may be taken as expressing scorn and disdain in another. Indeed, within our own culture the very same physical gesture could count as greeting someone, indicating your presence in a roll call, bidding at an auction, signaling for a left turn, and any number of other things. The factors that determine exactly what it is that you are doing when you produce a physical gesture include the customs, habits, and conventions that are in force as well as the particular circumstances at the time—a complex network of entrenched customs and practices, the agent's beliefs and intentions, her social relationships to other agents involved, and numerous other factors.

Considerations like these make it seem exceedingly unlikely that anyone could ever produce correct behavioral definitions of mental terms linking every mental expression with an equivalent behavioral expression referring solely to pure physical behavior ("motions and noises"). In fact, we have seen here how futile it would be to look for interesting generalizations, much less definitions, connecting mental states, like wanting to greet someone, with

physical behavior. To have even a glimmer of success, we would need, it seems, to work at the level of intentional action, not of physical behavior—that is, at the level of actions like greeting a friend, buying and selling, and reading the morning paper, not behavior at the level of motions and noises.

Do Pains Entail Pain Behavior?

Nevertheless, as noted earlier, some mental phenomena seem more closely tied to physical behavior—occurrences like pains and itches that have "natural expressions" in behavior. When you experience pain, you wince and groan and try to get away from the source of the pain; when you itch, you scratch. This perhaps is what gives substance to the talk of "pain behavior"; it is probably easier to recognize pain behavior than, say, greeting behavior, in an alien culture. We sometimes try to hide our pains and may successfully suppress winces and groans; nonetheless, pains do seem, under normal conditions, to manifest themselves in a roughly identifiable range of physical behavior. Does this mean that pains entail certain specific types of physical behavior?

Let us first get clear about what "entailment" is to mean in a context of this kind. When we say that pain "entails" winces and groans, we are saying that "Anyone in pain winces and groans" is *analytically*, or *conceptually*, *true*—that is, like "Bachelors are unmarried" and "Vixens are females," it is true *solely in virtue of the meanings of the terms involved* (or *the concepts expressed by these terms*). If "toothache" is definable, as Hempel claims, in terms of "weeping" and "making gesture G" (where we leave it to Hempel to specify G), toothache entails weeping and making gesture G in our sense. And if pain entails winces and groans, no organism could count as "being in pain" unless it could evince wincing and groaning behavior. That is, there is no "possible world" in which something is in pain but does not wince and groan.[16]

Some philosophers have argued that there is no pain-behavior entailment because pain behavior can be completely and thoroughly suppressed by some people, the "super-Stoics" and "super-Spartans" who have trained themselves not to show their pains in overt behavior.[17] This objection can be met, at least partially, by pointing out that super-Spartans, although they do not actually exhibit pain behavior, can still be said to have a *propensity*, or *disposition*, to exhibit pain behavior—that is, they *would* exhibit overt pain behavior *if certain conditions were to obtain* (for example, the super-Spartan code of conduct is renounced, their inhibition is loosened by alcohol, etc.). It is only that these conditions do not obtain for them, and so their behavior dispositions

associated with pain remain unmanifested. And a truthful super-Spartan will say yes when asked "Are you in pain?" although she will not groan, wince, or complain. There is this difference, then, between a super-Spartan who is in pain and another super-Spartan who is not in pain: It is true of the former, but not the latter, that if certain conditions were to obtain for her, she would exhibit pain behavior. It seems, therefore, that the objection based on the conceivability of super-Spartans can be substantially mitigated by formulating the entailment claim in terms of behavior dispositions or propensities rather than actual behavior production. After all, most behaviorists identify mentality with behavior dispositions, not actual behaviors.[18]

So the modified entailment thesis says this: It is an analytic, conceptual truth that anyone in pain has a propensity to wince or groan. Is this true? Consider animals: Dogs and cats can surely feel pain. Do they wince or groan? Perhaps. How about squirrels or bats? How about snakes and octopuses? Evidently, in order to groan or wince or emit a specified type of behavior (such as screaming and writhing in pain), an organism needs a certain sort of body and bodily organs with specific capacities and powers. Only animals with vocal cords can groan or scream; we can be certain that no one has ever observed a groaning snake or octopus! Thus, the entailment thesis under consideration has the consequence that organisms without vocal cords cannot be in pain, which is absurd. The point can be generalized: Whatever behavior type is picked, we can coherently imagine a pain-capable organism that is physically unsuited to produce behavior of that type.[19]

If this is the case, there is no specific behavior type that is entailed by pain. More generally, the same line of consideration should show that no specific behavior type is entailed by any mental state. And yet a weaker thesis, perhaps something like the following, may be true:

> *Weak Behavior Entailment Thesis*. For any pain-capable species[20] there is a certain behavior type B such that, for that species, being in pain entails a propensity to emit behavior of type B.

According to this thesis, then, each species may have its own special way of expressing pain behaviorally, although there are no universal and species-independent pain-to-behavior entailments. If this is correct, the concept of pain involves the concept of behavior only in this sense: Any organism in pain has a propensity to behave in some characteristic way. Note that the Weak Entailment Thesis is formulated in terms of a "propensity" to exhibit a type of behavior; having a propensity should be taken to mean that only when an appropriate

set of conditions obtains, the phenomenon will occur, or, alternatively, that there is a fairly high probability of its occurring. In any case, it is clear that there is no behavior pattern that can count as "pain behavior" across all pain-capable organisms (and perhaps also inorganic systems). Again, this makes the prospect of defining pain in terms of behavior exceedingly remote.

Ontological Behaviorism

Logical behaviorism is a thesis about the meanings of psychological expressions; as you recall, the claim is that the meaning of every psychological term is definable exclusively on the basis of behavioral-physical terms. More concretely, the claim is that given any sentence including psychological expressions, we can in principle produce a synonymous sentence devoid of psychological expressions. But we can also consider a behaviorist thesis about psychological states or phenomena as such, independently of the language in which they are described. The question—the "ontological" question—is what mental states *are*. Given that psychological sentences are translatable into behavioral sentences, does this mean that there are only behaviors, but no mental states? No pains; only pain behavior?

A radical behaviorist may claim that there are no mental facts over and above actual and possible behavioral facts, and that inner mental events do not exist, and that if they did, they are of no consequence. This is ontological behaviorism: Existentially, our mentality consists solely in behaviors and behavioral dispositions; there is nothing more. This, therefore, is a form of psychological eliminativism,[21] the view that mentality as ordinarily conceived is as misguided and defunct as the phlogiston theory of combustion and the neo-vitalist theory of entelechies as the "principle of life." Like such discredited scientific theory, mentalistic psychology will be jettisoned sooner and later. Such is the claim of radical behaviorism.

Compare the following two claims about pain:

1. Pain = winces and groans.
2. Pain = the cause of winces and groans.

Claim (1) expresses an ontological behaviorism about pain; it tells us what pain is—it is winces and groans. There is nothing more to pain than pain behavior—if there is also some private event going on, that is not pain, or part of pain, whatever it is, and it is psychologically irrelevant. But (2) is not a form of ontological behaviorism, since the *cause* of winces and groans need not be,

and probably isn't, more behavior. Clearly (2) may be affirmed by someone who thinks that it is an *internal state* of organisms (say, a neural state) that causes pain behavior, like winces and groans. Moreover, a dualist—even a Cartesian dualist—can welcome (2): She would say that a private mental event, an inner pain experience, is the cause of winces, groans, and other pain behavior. Further, we might even claim that (2) is analytically or conceptually true: The concept of pain is that of an internal state apt to cause characteristic pain behaviors like winces and groans.[22] Note this paradoxical result: (2) can be taken as an unexpected vindication of logical behaviorism about "pain," since it allows us to translate any sentence of the form "X is in pain" into "X is in a state that causes winces and groans," a sentence devoid of psychological expressions.[23] The same goes for other sentences including the term "pain." On the ontological question "What is pain really?" (2) is consistent with physicalism, property dualism, and even Cartesian interactionist dualism, though not with epiphenomenalism or Leibniz's preestablished harmony. One lesson of all this is that logical behaviorism does not entail ontological behaviorism.

Does ontological behaviorism entail logical behaviorism? Again, the answer has to be no. From the fact that Xs are Ys, nothing interesting follows about the meanings of the expressions "X" and "Y"—in particular, nothing follows about their interdefinability. Consider some examples: We know that bolts of lightning are electric discharges in the atmosphere and that genes are DNA molecules. But the expressions "lightning" and "electric discharge in the atmosphere" are not conceptually related, much less synonymous; nor are the expressions "gene" and "DNA molecule." So it may be that pain = winces, groans, and avoidance behavior. But you would not be able to verify that "pain" means the same as "winces, groans, and avoidance behavior" by consulting the most comprehensive dictionary.

In a similar vein, one could say, as some philosophers have argued,[24] that there are in this world no inner private episodes like pains, itches, and twinges but only observable behaviors or dispositions to exhibit such behaviors. One may say this because one holds a certain form of logical behaviorism or takes a dim view of supposedly private and subjective episodes in an inner theater. But one may be an ontological behaviorist on a methodological ground, affirming that there is *no need to posit* private inner events like pains and itches since they are not needed—nor are they able—to explain observed behaviors of humans and other organisms, as neural-physical states are sufficient for this purpose. A person holding such a view may well concede that the phenomenon purportedly designated by "pain" is an inner subjective state but will insist that there is no reason to think that the word actually refers to anything real (compare with

"witch," or "Bigfoot"). Daniel Dennett has urged that our concept of a private qualitative state ("qualia") is saddled with conditions that cannot be simultaneously satisfied, and that, as a result, there can be nothing that corresponds to the traditional idea of a private inner episode.[25] Paul Churchland and Stephen Stich have argued that beliefs, desires, and other intentional states as conceived in "folk" psychology will go the way of phlogiston and entelechies as systematic, scientific psychology makes progress.[26]

The Real Relationship Between Pain and Pain Behavior

Our discussion has revealed serious difficulties with any entailment claims about the relationship between pain and pain behavior, or more generally, between types of mental states and types of behavior. The considerations seemed to show that though our pains may cause our pain behaviors, this causal relation is a contingent fact. But leaving the matter there is unsatisfying: Surely pain behaviors—groans, winces, screams, writhings, attempts to get away, and such—have something important to do with our notion of pain. How else could we learn, and teach, the concept of pain or the meaning of the word "pain"? Wouldn't we rightly deny the concept of pain to a person who does not at all appreciate the connection between pains and these characteristic pain behaviors? If a person observes someone writhing on the floor, clutching his broken leg, and screaming for help and yet refuses to acknowledge that he is in pain, wouldn't it be correct to say that this person does not have the concept of pain, that he does not know what "pain" means? If Wittgenstein's "beetle in the box" shows anything, it is the point that publicly accessible behaviors are essential to anchor the meanings of our mental terms, like "pain," and explain the possibility of knowledge of what goes on in other people's minds. That is, observable behavior seems to have an essential grounding role for the semantics of our psychological language and the epistemology of other minds. What we need, therefore, is a positive account of the relationship between pain and pain behavior that explains their intimate connection without making it into one of logical, or conceptual, entailment.

The following is one possible story. Let us begin with an analogy: How do we fix the meaning of "one meter long"—that is, the concept of a meter? We sketch an answer based on Saul Kripke's influential work on names and their references.[27] Consider the Standard Meter: a bar of platinum-iridium alloy kept in a vault near Paris.[28] Is the following statement necessarily, or analytically, true?

The Standard Meter is one meter long.

There is a clear sense in which the Standard Meter *defines* what it is to be one meter in length. But does being the Standard Meter (or having the same length as the Standard Meter) entail being one meter long? The Standard Meter is a particular physical object, manufactured at a particular date and place and now located somewhere in France, and surely this metallic object might not have been the Standard Meter and might not have been one meter long. (It could have been fashioned into a bowl, or it could have been made into a longer rod of two meters.) In other words, it is a contingent fact that this particular platinum-iridium rod was selected as the Standard Meter, and it is a contingent fact that it is one meter long. No middle-sized physical object has the length it has necessarily; anything could be longer or shorter—or so it seems. We must conclude, then, that the statement that something has the same length as the Standard Meter does not logically entail that it is one meter long, and it is not analytically, or conceptually, true that if the length of an object coincides with that of the Standard Meter, it is one meter long.

But what, then, is the relationship between the Standard Meter and the concept of the meter? After all, the Standard Meter is not called that for nothing; there must be some intimate connection between the two. A plausible answer is that we specify the property of being one meter long (the meaning, if you wish, of the expression "one meter long") by the use of a *contingent* relationship in which the property stands. One meter is the length of this thing (namely, the Standard Meter) here and now. It is only contingently one meter long, but that is no barrier to using it to specify what counts as one meter. This is just like when we point to a ripe tomato and say, "Red is the color of this tomato." It is only a contingent fact that this tomato is red (it could have been green), but we can use this contingent fact to specify what the color red is and what the word "red" means.

Let us see how a similar account might go for pain: We specify what pain is (or fix the meaning of "pain") by reference to a contingent fact about pain, namely, that pain causes winces and groans in humans. This is a contingent fact about this world. In worlds in which different laws hold, or worlds in which the central nervous systems of humans and those of other organisms are hooked up differently to peripheral sensory surfaces and motor output systems, the patterns of causal relations involving pain may be very different. But as things stand in this world, pain is the cause of winces and groans and certain other behaviors in humans and related animal species. In worlds in which pains do not cause winces and groans, different behaviors may count

as pain behavior, in which case pain specifications in those worlds could advert to the behaviors caused by pains there. This is similar to the color case: If cucumbers but not ripe tomatoes were red, we would be specifying what "red" means by pointing to cucumbers instead.

The foregoing is only a sketch of an account but not an implausible one. It explains how (2) above ("Pain = the cause of winces and groans"), though only contingently true, can help specify what pain is and fix the reference of the term "pain." And it seems to show a good fit with the way we learn, and teach, how to use the word "pain" and other mental expressions denoting sensations. The approach brings mental expressions under the same rubric with many other expressions, as we have seen, such as "red" and "one meter long." Though not implausible, the story may not be over just yet: The reader is encouraged to think about the possible differences between the case of pain and cases like the color red and one meter in length. In particular, think about how it deals with, or fails to deal with, the conundrum of Wittgenstein's "beetle in the box."

Behaviorism in Psychology

So far we have been discussing behaviorism as a philosophical doctrine concerning the meanings of mental terms and the nature of mental states. But as we noted at the outset, "behaviorism" is also the name of an important and influential psychological movement initiated early in the twentieth century that came to dominate scientific psychology and the social sciences in North America and many other parts of the world for several decades. It held its position as the reigning methodology of the "behavioral sciences" until the latter half of the century, when "cognitivism" and "mentalism" began a strong comeback and replaced it as the new orthodoxy.

Behaviorism in science can be viewed in two ways: First, as a precept on how psychology should be conducted as a science, it provides guidance to questions like what its proper domain should be, what conditions should be placed on admissible evidence, what its theories are supposed to accomplish, by what standards its explanations are to be evaluated, and so on. Second, behaviorism, especially B. F. Skinner's "radical behaviorism," is a specific behaviorist research paradigm seeking to construct psychological theories conforming to a fairly explicit and precisely formulated pattern (for example, Skinner's "operant conditioning"). Here we have room only for a brief and sketchy discussion of scientific behaviorism in the first sense. Discussion of Skinner's radical behaviorism is beyond the scope of this book.

We can begin with what may be called methodological behaviorism:

(I) The only admissible evidence for the science of psychology is observable behavioral data—that is, data concerning the observable physical behavior of organisms.

We can understand (I) somewhat more broadly than merely as a stricture on admissible "evidence" by focusing on the "data" it refers to. Data serve two closely related purposes in science: First, they constitute the domain of phenomena for which theories are constructed to provide explanations and predictions; second, they serve as the evidential basis that can support or undermine theories. What (I) says, therefore, is that psychological theories should attempt to explain and predict only data concerning observable behavior and that only such data should be used as evidence against which psychological theories are to be evaluated. These two points can be seen to collapse into one when we realize that explanatory and predictive successes and failures constitute, by and large, the only measure by which we evaluate how well theories are supported by evidence.

The main reason some psychologists and philosophers have insisted on the observability of psychological data is to ensure the *objective* or *intersubjective testability* of psychological theories. It is thought that introspective data—data obtained by a subject by inwardly inspecting her own inner Cartesian theater—are essentially private and subjective and hence cannot serve as the basis for intersubjective validation of psychological theories. In short, the idea is that intersubjective access to data is required to ensure the possibility of intersubjective agreement in science and that the possibility of intersubjective agreement is required to ensure the objectivity of psychology. Only behavioral (and more broadly, physical) data, it is thought, meet the condition of intersubjective observability. In short, (I) aims at securing the objectivity of psychology as a science.

What about a subject's verbal reports of her inner experiences? A subject in an experiment involving mental imagery might report: "I am now rotating the figure counterclockwise." What is wrong with taking the following as an item of our data: Subject S is rotating her mental image counterclockwise? Someone who holds (I) will say something like this: Strictly speaking, what we can properly consider an item of data here is S's utterance of the words "I am now rotating the figure counterclockwise." Counting S's actual mental operation of rotating her mental image as a datum involves the assumption that she is a

competent speaker of English, that intersubjective meaning can be attached to reports of inner experience, and that she is reporting her experience correctly. These are all substantial psychological assumptions, and we cannot consider the subject's reports of her visual activity to meet the criterion of intersubjective verifiability. Therefore, unless these assumptions themselves can be behaviorally justified, the cognitive scientist is entitled only to the subject's utterance of the string of words, not the presumed content of those words, as part of her basic data.

Consciousness is usually thought to fall outside the province of psychological explanation for the behaviorist. Inner conscious states are not among the phenomena it is the business of psychological theory to explain or predict. In any case, many psychologists and cognitive scientists may find (I) by and large acceptable, although they are likely to disagree about just what is to count as *observable* behavior. (Some may consider verbal reports, with their associated meanings, as admissible data, especially when they are corroborated by nonverbal behavior.)

A real disagreement arises, though, concerning the following stronger version of methodological behaviorism:

> (II) Psychological *theories* must not invoke the *internal states* of psychological subjects; that is, psychological explanations must not appeal to internal states of organisms, nor should references to such states occur in deriving predictions about behavior.

This appears to have been a tenet of Skinner's psychological program. On this principle, organisms are to be construed as veritable black boxes whose internal structure is forever closed to the psychological investigator. Psychological generalizations, therefore, must only correlate observable stimulus conditions as input, behavioral outputs, and subsequent reinforcements. But isn't it obvious that when the same stimulus is applied to two organisms, they can respond with different behavior output? How can we explain behavioral differences elicited by the same stimulus condition without invoking differences in their internal states?

The Skinnerian answer is that such behavioral differences can be explained by reference to the differences in the *history* of reinforcement for the two organisms; that is to say, the two organisms emit different behavior in response to the same stimulus because their *histories* involving external stimuli, elicited behaviors, and the reinforcements following the behaviors are different. But if such an explanation works, isn't that because the differences in the

histories of the two organisms led to differences in their present internal states? Isn't it plausible to suppose that these differences *here and now* are what is directly implicated in the production of different behaviors *now*? To suppose otherwise would be to embrace "mnemic" causation—causal influence that leaps over a temporal gap with no intermediate links bridging cause and effect. Apart from such metaphysical doubts, there appears to be an overwhelming consensus at this point that the stimulus-response-reinforcement model is simply inadequate to generate explanatory or predictive theories for vast areas of human and animal behaviors.

And why is it impermissible to invoke present internal differences as well as differences in histories to explain differences in behavior output? Notice how sweeping the constraint expressed by (II) really is: It outlaws references not only to inner mental states of the subject but also to its internal physical-biological states. Methodological concerns with the objectivity of psychology as a science provide an intelligible (if perhaps not sufficient) motivation for banishing the former, but it seems clearly insufficient to justify banning the latter from psychological theories and explanations. Even if it is true, as Skinner claims,[29] that invoking internal neurobiological states does not help psychological theorizing, that hardly constitutes a sufficient ground for prohibiting it as a matter of scientific methodology.

In view of this, we may consider a further version of behaviorism as a rule of psychological methodology:

> (III) Psychological theories must make no reference to inner *mental* states in formulating psychological explanations.

This principle allows the introduction of internal biological-physical states, including states of the central nervous system, into psychological theories and explanations, prohibiting only reference to inner mental states. But what is to count as such a state? Does this principle permit the use of such concepts as "drive," "information," "memory," "attention," "mental representation," and the like in psychological theories? To answer this question we would have to examine these concepts in the context of particular psychological theories making use of them; this is not a task for armchair philosophical conceptual analysis. We should keep in mind, though, that the chief rationale for (III)—in fact, the driving motivation for the entire behaviorist methodology—is the insistence on the objective testability of theories and public access to sharable data. This means that what (III) is intended to prohibit is the introduction of *private subjective* states for which objective access is thought to be problematic, not the use

of theoretical constructs posited by psychological theories for explanatory and predictive purposes, as long as these meet the requirement of intersubjectivity. Unlike overt behavior, these constructs are not, as a rule, "directly observable," and they are not strictly definable or otherwise reducible in terms of observable behavior. However, they differ from the paradigmatic inner mental states in that they apparently do not show the first-person/third-person asymmetry of epistemic access. Scientific theories often introduce theoretical concepts for entities (electrons, magnetic fields, quarks) and properties (spin, polarization) that go far beyond the limits of human observation. Like any other science, psychological theory should be entitled to such theoretical constructs.

But in excluding private conscious states from psychological theory, (III) excludes them from playing any causal-explanatory role in relation to behavior. If it is true, as we ordinarily think, that some of our behavior is caused by inner mental states disallowed by (III), our psychological theory is likely to be incomplete: There may well be behavior for which no theory meeting (III) can provide full explanations. (Some of these issues are discussed further in chapters 7, 9, and 10.)

Are there other methodological constraints for psychological theory? How can we be sure that the states and entities posited by a psychological theory (for example, "intelligence," "mental representation," "drive reduction") are "real"? If, in explaining the same data, one psychological theory posits one set of unobservable states and another theory posits an entirely different set, which theory, if any, should be believed? That is, which theory represents the *psychological reality* of the subjects? Does it make sense to raise such questions? If it does, should there be the further requirement that the entities and states posited by a psychological theory have a "biological reality"—that is, must they somehow be "realized" or "implemented" in the biological-physical structures and processes of the organism? These are important questions about the science of psychology, and we deal with some of them later in our discussion of mind-body theories and the status of cognitive science (chapters 5, 6).

Why Behavior Matters to Mind

Our discussion thus far has been, by and large, negative toward behaviorism. This should not be taken to mean that we should take a negative attitude on the relevance of behavior to minds. The fact is that the importance of behavior to mentality cannot be overemphasized. In retrospect, it seems that, impressed by the crucial role of behavior to mentality, various forms of behaviorism, in

particular logical behaviorism, got carried away, going way overboard and advocating extreme and unrealistic theories, with a reformer's zeal.

There are three main players on the scene in discussions of mentality: mind, brain, and behavior. An important task of the mind-body problem is to elucidate the relationships among these three elements. The detailed issues and problems are yet to be discussed in the rest of this book. But here is a rough picture:

1. The brain is the ontological—that is, existential—base of the mind.
2. The brain, and perhaps the mind also, is the cause of behavior.
3. Behavior is the semantic foundation of mental language. It is what fixes the meanings of our mental/psychological expressions.
4. Behavior is the primary, almost exclusive, evidence for the attribution of mental states to other beings with minds. Our knowledge of other minds depends primarily on observation of behavior.

It is fair to say these statements are what most of us believe. There will be dissenters, especially about (1) and (2)—for example, Cartesian dualists. What concerns us here are items (3) and (4). Without behavior, it is hard to see how our mental terms can acquire their common, public meanings fit for interpersonal communication. And without behavioral evidence (including verbal behavior), it is not possible to know what others are thinking and feeling. (Try to imagine how you might find out what an immaterial soul is thinking or feeling.) If we were to lose observational access to others' behavior, the fabric of our social relationships would completely unravel. Unquestionably, behavior is the semantic and epistemological foundation of our mental and social life.

To summarize, the brain is what existentially underlies, and supports, our mental life. You take away the brain, and mental life is no more. Behavior, on the other hand, is the semantical and epistemological foundation of mentality. Without it, psychological language would be impossible, and we could never know what goes in other minds. It is impossible to exaggerate, or even underplay, the crucial place observable behavior has in our social life.

For Further Reading

The influential classic work representing logical behaviorism is Gilbert Ryle, *The Concept of Mind*. Also important are Rudolf Carnap, "Psychology in Physical Language," and Carl G. Hempel, "The Logical Analysis of Psychology."

For an accessible Wittgensteinian perspective on mind and behavior, see Norman Malcolm's contributions in *Consciousness and Causality* by D. M. Armstrong and Norman Malcolm. For scientific behaviorism, see B. F. Skinner's *Science and Human Behavior* and *About Behaviorism*. Both are intended for nonspecialists.

For a historically important critique of Skinnerian behaviorism, see Noam Chomsky's review of Skinner's *Verbal Behavior*. For criticism of logical behaviorism, see Roderick M. Chisholm, *Perceiving*, pp. 173–185; and Hilary Putnam, "Brains and Behavior." George Graham's article "Behaviorism" in the *Stanford Encyclopedia of Philosophy* is a useful resource; so is Georges Rey's entry "Behaviorism" in the *Macmillan Encyclopedia of Philosophy*, 2nd ed.

Notes

1. William James, *The Principles of Psychology*, p. 15. Page references are to the 1981 edition.

2. Ibid., p. 185.

3. J. B. Watson, "Psychology as the Behaviorist Views It," p. 158.

4. William James, *The Principles of Psychology*, p. 21 (emphasis in original).

5. This piquant term comes from Daniel C. Dennett, *Consciousness Explained*. Dennett considers the Cartesian theater an incoherent myth.

6. Ludwig Wittgenstein, *Philosophical Investigations*, section 293. We need to assume that there are no beetles flying around for everyone to see!

7. For the notion of behavior as internally caused bodily motion, see Fred Dretske, *Explaining Behavior*, chapters 1 and 2.

8. You might feel uncomfortable about the last two examples: Perhaps our bodies do these things, but it sounds odd to say that *we* do these things. The ordinary notion of doing seems to involve the idea of voluntariness; however, the notion of behavior appropriate to behaviorism need not include such an element.

9. Carl G. Hempel, "The Logical Analysis of Psychology," p. 91.

10. Positivists, including Hempel, often used a much looser sense of definition (and translatability); however, for logical behaviorism to be a significant thesis, we need to construe definition in a more strict sense.

11. Carl Hempel, "The Logical Analysis of Psychology," p. 17.

12. There is a long-standing controversy in moral theory as to whether certain beliefs (for example, the belief that you have a moral duty to help a friend), without any associated desires, can motivate a person to act. The dispute, however, concerns only a small class of beliefs, chiefly evaluative and

normative beliefs about what ought to be done, what is desirable, and the like. The view that to generate an action both desire and belief must be present is usually attributed to Hume.

13. For an early statement of this point, see Roderick M. Chisholm, *Perceiving*.

14. This may be what is distinctive and interesting about the "ceteris paribus" clauses qualifying psychological generalizations—in particular, those concerning motivation and action.

15. The phenomena discussed in this paragraph and the next are noted in Berent Enç, "Redundancy, Degeneracy, and Deviance in Action."

16. Strictly speaking, this last sentence defines "metaphysical" entailment, as distinguished from analytical or conceptual entailment as defined earlier. There are differences between them that can be important in some context; however, this will not affect our discussion.

17. Hilary Putnam, "Brains and Behavior."

18. Moreover, many mental states have bodily manifestations; pain may be accompanied by a rise in blood pressure and a quickening pulse, and super-Spartans presumably could not "hide" these physiological signs of pain (recall Hempel's behavioral translation of "Paul has a toothache"). Whether these count as "behavior" may only be a verbal issue in this context.

19. Perhaps this can be called a "multiple realizability" thesis in regard to behavior. On multiple realizability, see chapter 5. Whether it has consequences for behaviorism that are similar to the supposed consequences of the multiple realizability of mental states is an interesting further question.

20. Species may be too wide here, given that expressions of pain are, at least to some extent, culture-specific and can even differ from person to person within the same culture.

21. See Paul Churchland, "Eliminative Materialism and the Propositional Attitudes."

22. See chapters 5 and 6 for discussion of the functionalist conception of pain as that of a "causal intermediary" between certain stimulus conditions (for example, tissue damage) and characteristic pain behaviors.

23. This does not mean that the original logical behaviorists, like Hempel and Gilbert Ryle, would have accepted (2) as a behavioral characterization of "pain." The point, however, is that it meets Hempel's translatability thesis—his form of logical behaviorism. Note that the "cause" is a topic-neutral term—it is neither mental nor behavioral-physical.

24. See Gilbert Ryle, *The Concept of Mind*.

25. Daniel Dennett, "Quining Qualia."

26. Paul Churchland, "Eliminative Materialism and the Propositional Attitudes"; Stephen Stich, *From Folk Psychology to Cognitive Science: The Case Against Belief*.

27. See Saul Kripke, *Naming and Necessity*.

28. The meter is no longer defined this way; the current definition, adopted in 1984 by the General Conference on Weights and Measures, is reportedly based on the distance traveled by light through a vacuum in a certain (very small) fraction of a second.

29. See B. F. Skinner, *Science and Human Behavior*.

CHAPTER 4

Mind as the Brain

The Psychoneural Identity Theory

Some ancient Greeks thought that the heart was the organ responsible for thoughts and feelings—an idea that has survived, we are told, in the traditional symbolism of the heart as signifying love and romance. But the Greeks got it wrong; we now know, as surely as such things can be known, that the brain is where the action is as far as our mental life is concerned. If you ask people where their minds or thoughts are located, they will point to their heads. Does this mean only that the mind and brain share the same location, or something stronger, namely, that the mind *is* the brain? We consider here a theory that advocates this stronger claim—that the mind is identical with the brain and that for a creature to have mentality is for it to have a brain with appropriate structure and capacities.

Mind-Brain Correlations

But what makes us think that the brain is "the seat of our mental life," as Descartes might have put it? The answer seems clear: There are *pervasive and systematic psychoneural correlations*, that is, *correlations between mental phenomena and neural states of the brain*. This is not something we know a priori; we know it from empirical evidence. We observe that injuries to the brain often have a dramatic impact on mental life, affecting the ability to reason, recall, and perceive, and that they can drastically impair a person's cognitive capacities and even alter her personality traits. Chemical changes in the brain brought on by ingestion of alcohol, antidepressants, and other psychoactive drugs affect our moods, emotions, and cognitive functions. When a brain concussion

knocks us out, our conscious life goes blank. Sophisticated brain imaging techniques allow us to "see" just what is going on in our brains when we are engaged in certain mental activities, like seeing green or feeling agitated. It is safe to say that we now have overwhelming scientific evidence attesting to the centrality of the brain and its activities as determinants of our mental life.

A badly scraped elbow can cause you a searing pain, and a mild food poisoning is often accompanied by stomachaches and queasy feelings. Irradiations of your retinas cause visual sensations, which in turn cause beliefs about objects and events around you. Stimulations of your sensory surfaces lead to sensory and perceptual experiences of various kinds. However, peripheral neural events are only remote causes; we think that they bring about conscious experiences only because they cause appropriate states of the brain. This is how anesthesia works: If the nerve signals coming from sensory peripheries are blocked or the normal functions of the brain are interfered with so that the central neural processes that underlie conscious experience are prevented from occurring, there will be no experience of pain—perhaps no experience of anything. It is plausible that everything that occurs in mental life has a state of the brain (or the central nervous system) as its *proximate* physical basis. It would be difficult to deny that the very existence of our mentality depends on the existence of appropriately functioning neural systems: If all the cells and molecules that make up your brain were scattered in intergalactic space, your whole mental life would vanish at that moment, just as surely as annihilating all the molecules making up your body would mean its end. At least that is the way things seem. We may summarize this in the following thesis:

> *Mind-Brain Correlation Thesis*. For each type M of mental event that occurs to an organism *o*, there exists a brain state of kind B (M's "neural correlate" or "substrate") such that M occurs to *o* at time *t* if and only if B occurs to *o* at *t*.

According to this thesis, then, each type of mental event that can occur to an organism has a neural correlate that is both necessary and sufficient for its occurrence. So for each organism there is a set of mind-brain correlations covering every kind of mental state it is capable of having.

Two points may be noted about these brain-mind correlations:

1. They are "lawlike": The fact that pain is experienced when certain of your neurons (say, C-fibers and Aδ-fibers) are activated is a matter of *lawful regularity*, not accidental, or coincidental, co-occurrences.

2. Even the smallest change in your mental life cannot occur unless there are some specific (perhaps still unknown) changes in your brain state; for example, when your headache goes away, there must be an appropriate change in your neural states.

Another way of putting these points, though this is not strictly equivalent, is to say that mentality *supervenes* on brain states. Remember that this supervenience, if it indeed holds, is something we know from observation and experience, not a priori. Moreover, specific correlations—that is, correlations between specific types of mental states (say, pain) and specific types of brain states (say, the activation of certain neural fibers)—are again matters of scientific research and discovery, and we may assume that many of the details about these correlations are still largely unknown. However, it is knowledge of these specific correlations, rough and incomplete though it may be, that ultimately underlies our confidence in the general thesis of mind-brain correlation and mind-brain supervenience. If Aristotle had been correct (and he *might* have been correct) about the heart being the engine of our mentality, we would have a mind-heart correlation thesis and mind-heart supervenience, instead of the mind-brain correlation thesis and mind-brain supervenience.

Making Sense of Mind-Brain Correlations

When a systematic correlation between two properties or types of events has been observed, we want an explanation, or interpretation, of the correlation: Why do the properties F and G correlate? Why is it that an event of type F occurs just when an event of type G occurs? We do not want to countenance too many "brute," unexplained coincidences in nature. An explanatory demand of this kind becomes even more pressing when we observe systematic patterns of correlation between two large families of properties, like mental and neural properties. Let us first look at some examples of property correlations outside the mind-brain case:

a. Whenever the ambient temperature falls below 20 degrees Fahrenheit and stays there for several days, the local lakes and ponds freeze over. Why? The answer, of course, is that the low temperature *causes* the water in the ponds to freeze. The two events are *causally related*, and that is why the observed correlation occurs.
b. You enter a clock shop and find an astounding scene: Dozens and dozens of clocks of all shapes and sizes are busily ticking away, and

they all show exactly the same time, 2:00. Awhile later, you see all of them showing exactly 2:30, and so on. What explains this marvelous correlation among these clocks? It could not be a coincidence, we think. One possible answer is that the shopkeeper synchronized all the clocks, which are all working properly, before the shop opened in the morning. Here, a *common cause*, the shopkeeper's action in the morning, explains the correlations that are now observed; to put it another way, one clock showing 3:30 and another showing the same time are *collateral effects of a common cause*. There are no direct causal relationships between the clocks that are responsible for the correlations.

c. We can imagine a slightly different explanation of why the clocks are keeping the same time: These clocks actually are not very accurate, and some of them gain or lose time markedly every five minutes or so. But there is a little leprechaun whose job is to run around the shop, unseen by the customers, synchronizing the clocks every minute. That is why every time you look, the clocks show the same time. This again is a *common-cause* explanation of a correlation, but it is different from the story in (b) in the following respect: This explanation involves a continued intervention of a causal agent, whereas in (b) a single cause in the past is sufficient. In neither case, however, is there a direct cause-effect relationship between the correlated events.

d. Why do temperature and pressure covary for gases confined in a rigid container? The temperature and pressure of a gas are both dependent on the motions of the molecules that compose the gas: The temperature is the average kinetic energy of the molecules, and the pressure is the momentum imparted to the walls of the container (per unit area) by the molecules colliding with them. Thus, the rise in temperature and the rise in pressure can be viewed as *two aspects* of one and the same underlying microprocess.

e. Why does lightning occur just when there is an electric discharge between clouds or between clouds and the ground? Because lightning simply *is* an electric discharge involving clouds and the ground. There is here only one phenomenon, not two that are correlated with each other, and what we thought were distinct correlated phenomena turn out to be one and the same event, under two different descriptions. Here an apparent correlation turns out to be an *identity*.

f. Why do the phases of the moon (full, half, quarter, and so on) covary with the tidal actions of the ocean (spring tides, neap tides, and so on)? Because the relative positions of the earth, the moon, and the sun determine both the phases of the moon and the combined strength of the gravitational forces of attraction exerted on the ocean water by the moon and the sun. So the changes in gravitational force are the proximate causes of tidal actions, and the relative positions of the three bodies can be thought of as their distal cause. The phases of the moon are merely collateral effects of the positions of the three bodies involved and serve only as an indication of what the positions are (full moon when the earth is between the sun and the moon on a straight line, and so on), having no causal role whatever on tidal actions.

What about explaining, or interpreting, mind-brain correlations? Which of the models we have surveyed best fits the mind-body case? As we would expect, all of these models have been tried. We begin with some causal approaches to the mind-body relation:

Causal Interactionism. Descartes thought that causal interaction between the mind and the body occurred in the pineal gland (chapter 2). He speculated that "animal spirits"—fluids made up of extremely fine particles flowing around the pineal gland—cause it to move in various ways, and these motions of the gland in turn cause conscious states of the mind. Conversely, the mind could cause the gland to move in various ways, affecting the flow of the surrounding animal spirits. This in turn influenced the flow of these fluids to different parts of the body, ultimately issuing in various physiological changes and bodily movements.[1]

"Preestablished Harmony" Between Mind and Body. Leibniz, like many of his great contemporary Rationalists, thought that no coherent sense could be made of Descartes's idea that an immaterial mind could causally influence, or be influenced by, a material body like the pineal gland, managing to move this not-so-insignificant lump of tissue hither and thither. On his view, the mind and the body are in a "preestablished harmony," rather like the clocks that were synchronized by the shopkeeper in the morning, with God having started off our minds and bodies in a harmonious relationship. Whether this is any less fantastical an idea, at least for us, than Descartes's idea of mind-body interaction is debatable.

Occasionalism. According to Nicolas Malebranche, another major Continental Rationalist, whenever a mental event appears to cause a physical event or a physical event appears to cause a mental event, it is only an illusion. There is no direct causal relation between "finite minds" and bodies; when a mental event, say, your will to raise your arm, occurs, that only serves as an *occasion* for God to intervene and cause your arm to rise. Divine intervention is also responsible for the apparent causation of mental events by physical events: When your finger is cut, that again is an occasion for God to step in and cause you pain. The role of God, then, is rather like that of the leprechaun in the clock shop whose job is to keep the clocks synchronized at all times by continuous interventions. This view is known as occasionalism; it was an outcome of the doctrine, accepted by Malebranche and many others at the time, that God is the only genuine causal agent in this world, and that the apparent causal relations we observe in the created world are only that, an appearance.

The Double-Aspect Theory. Spinoza, another great Rationalist of the time, maintained that mind and body are simply two correlated aspects of a single underlying substance that is in itself neither mental nor material. This theory, like the doctrine of preestablished harmony and occasionalism, denies direct causal relationships between the mental and the physical; however, unlike them, it does not invoke God's causal action to explain the mental-physical correlations. The observed correlations are there because they are two distinguishable aspects of one underlying reality. A modern form of this approach is known as neutral monism, according to which the fundamental reality is neutral in the sense that it is intrinsically neither physical nor mental.

Epiphenomenalism. According to T. H. Huxley, a noted British biologist of the nineteenth century, all conscious events are caused by neural events in the brain, but they have no causal power of their own, being the ultimate end points of causal chains.[2] So all mental events are effects of the physiological processes in the brain, but they are powerless to cause anything else—even other mental events. You "will" your arm to rise, and it rises. But to think that your volition is the cause of the rising of the arm is to commit the same error as thinking that the changes in the phases of the moon cause the changes in tidal motions. The real cause of the arm's rising is a certain neural event in your brain, and this event also causes your experience of a volition to raise the arm. This is like the case of the moon and the tides: The relative positions of the earth, the moon, and the sun are the true cause of both the tidal motions

and the phases of the moon. Many scientists in brain research seem to hold, at least implicitly, a view of this kind (see chapter 10).

Emergentism. There is another interesting response to the question "Why are mental phenomena correlated with neural phenomena in the way they are?" It is this: The question is unanswerable—the correlations are "brute facts" that we must simply accept; they are not subject to further explanation. This is the position of emergentism. It holds that when biological processes attain a certain level of organizational complexity, a wholly new type of phenomenon, namely, consciousness and rationality, "emerges," and why and how these phenomena emerge is not explainable in terms of the lower-level physical-biological facts. There is no explanation of why, say, pains rather than itches emerge from C-fiber activations or why pains emerge from C-fiber activations rather than another type of neural state. That there are just these emergence relationships and not others must be accepted, in the words of Samuel Alexander, a leading theoretician of the emergence school, "with natural piety."[3] The phenomenon of emergence must be recognized as a fundamental fact about the natural world. One important difference between emergentism and epiphenomenalism is that the former, but not the latter, acknowledges causal power and efficacy of emergent mental phenomena.

The Psychoneural (or Psychophysical, Mind-Body) Identity Theory. This position, explicitly advanced as a solution to the mind-body problem in the late 1950s, advocates the *identification* of mental states with the physical processes in the brain. Just as there are no bolts of lightning *over and above* atmospheric electrical discharges, there are no mental events *over and above*, or *in addition to*, the neural processes in the brain. "Lightning" and "electrical discharge" are not dictionary synonyms, and the Greeks probably knew something about lightning but nothing about electric discharges; nonetheless, bolts of lightning are just electric discharges, and the two expressions "lightning" and "atmospheric electric discharge" refer to the same phenomenon. In the same way, the terms "pain" and "C-fiber activation" do not have the same dictionary meaning; Socrates knew a lot about pains but nothing about C-fiber stimulation. And yet pains turn out to be the activations of C-fibers, just as bolts of lightning turned out to be electrical discharges. In many ways, mind-brain identity seems like a natural position to take; it is not just that we point to our heads when we are asked where our minds are. Unless you are prepared to embrace Cartesian immaterial mental substances outside physical space, what

could your mind be if not your brain? And what could mental states be if not states of the brain?

* * *

But what are the arguments that support the identification of mental events with brain events? Even if your mind is in your head, your mind and your brain might only share the same space while remaining distinct. So are there good reasons for thinking that the mind *is* the brain? There are three principal arguments for the mind-brain identity theory. These are the simplicity argument, the explanatory argument, and the causal argument. We will see how these arguments can be formulated and defended, and try to assess their cogency. We will then turn to some arguments designed to refute, or at least discredit, the mind-brain identity theory.

The Argument from Simplicity

J. J. C. Smart, whose 1959 essay "Sensations and Brain Processes" had a critical role in establishing the psychoneural identity theory as a major position on the mind-body problem, emphasized the importance of *simplicity* as a ground for accepting the theory.[4] He writes:

> Why do I wish [to identify sensations with brain processes]? Mainly because of Occam's razor. . . . There does seem to be, so far as science is concerned, nothing in the world but increasingly complex arrangements of physical constituents. All except for one place: in consciousness. That is, for a full description of what is going on in a man you would have to mention not only the physical processes in his tissues, glands, nervous system, and so forth, but also his states of consciousness: his visual, auditory, and tactual sensations, his aches and pains. That these should be *correlated* with brain processes does not help, for to say that they are *correlated* is to say that they are something "over and above." . . . So sensations, states of consciousness, do seem to be the one sort of thing left outside the physicalist picture, and for various reasons I just cannot believe that this can be so. That everything be explicable in terms of physics . . . except the occurrence of sensations seems to me frankly unbelievable.[5]

Occam's (or Ockham's) razor, named after the fourteenth-century philosopher William of Ockham, is a principle that urges simplicity as an important

virtue of theories and hypotheses. The following two formulations are among the standard ways of stating this principle:[6]

I. Entities must not be multiplied beyond necessity.
II. What can be done with fewer assumptions should not be done with more.

Principle (I) urges us to adopt the simplest ontology possible, one that posits no unnecessary entities—that is, entities that have no work to do. In mathematics, we deal with natural numbers, rationals, and reals. But real numbers can be constructed out of rationals, which in turn can be constructed out of natural numbers. Natural numbers, too, can be generated as a series of sets. Sets are all we need to do mathematics. A crucial question in applying this principle, of course, is to determine what counts as going "beyond necessity," or what "work" needs to be done. The physicalist would hold that Cartesian immaterial minds are useless and unneeded posits; the Cartesian dualist, however, would disagree precisely on that point.

Principle (II) can be taken as urging simplicity and economy in theory construction: Choose the theory that gives the simplest, most parsimonious descriptions and explanations of the phenomena in its domain—that is, the theory that does its work with the fewest independent hypotheses and assumptions. When Napoleon asked the astronomer and mathematician Pierre de Laplace why God was absent from his theory of the planetary system, Laplace is reported to have replied, "Sir, I have no need of that hypothesis." To explain what needs to be explained (the stability of the planetary system, in this instance), we do well enough with physical laws alone; we need no help, and get none, from the "hypothesis" that God exists. Here, he is invoking version (II) of Ockham's razor. We can also see Laplace as invoking version (I): We don't need God in our ontology to do planetary astronomy; he would be an idler with no work to do.

There seem to be three lines of consideration one might pursue in attempting to argue in favor of the mind-brain identity theory on the ground of simplicity.

First, it is a simple fact that identification reduces the number of putative entities and thereby enhances ontological simplicity. When you say *X* is the same thing as *Y*—or, as Smart puts it, that *X* is nothing "over and above" *Y*—you are saying that there is just one thing here, not two. So if pain as a mental kind is identified with its neural correlate, we simplify our ontology on two levels: First, there is no mental kind, being in pain, in addition to C-fiber

stimulation; second—and this follows from the previous point—there are no individual pain occurrences in addition to occurrences of C-fiber stimulation. In this rather obvious way, mind-brain identification simplifies our ontology.

Second, it may also be argued that psychoneural identification is conducive to conceptual or linguistic simplicity as well. If all mental states are systematically identified with their neural correlates, there is a sense in which mentalistic language—language in which we speak of sensations, emotions, and thoughts—is *in principle* replaceable by a physical language in which we speak of neural processes. The mentalistic language is practically indispensable and we can be certain that it will remain so. We will almost certainly never have a full catalog of mental-neural correlations, and who among us will want to learn the bewilderingly complex and arcane medical terms? Still, we cannot deny the following crucial fact: On the identity theory, descriptions formulated in a mental vocabulary do not report facts or phenomena distinct from those reportable by sentences in a comprehensive physical-biological language. There are no excess facts beyond physical facts that can only be described in some nonphysical language. In this sense, physical language would be complete and universal.

Third, and this is what Smart seems to have in mind, suppose we stop short of identifying pain with C-fiber stimulation and stick with the correlation "Pain occurs if and only if (iff) Cfs occurs." As earlier noted, correlations cry out for explanation. How might such correlations be explained? In science, we standardly explain laws and correlations by deriving them from other, more fundamental laws and correlations. From what more basic correlations could we derive "Pain occurs iff Cfs occurs"? It seems quite certain that it cannot be derived from purely physical-biological laws alone. The simple reason is that these laws do not even speak of pain; the term, or concept, "pain" does not appear in physical-biological laws, for the obvious reason that it is not part of the physical-biological language. So if the pain-Cfs correlation is to be explained, its explanatory premises (premises from which it is to be derived) will have to include at least one law correlating some mental phenomenon with a physical-biological phenomenon—that is, at least one psychoneural correlation. But this puts us back in square one: How do we explain this perhaps more fundamental mental-physical correlation?

The upshot is that we are likely to be stuck with the pain-Cfs correlation and countless other such psychoneural correlations, one for each distinct type of mental state. (Think about how many mental states there are or could be, and in particular, consider this: For each declarative sentence *p*, such as "It will

snow tomorrow," there is the belief that *p*—that is, the belief that it will snow tomorrow.) And all such correlations would have to be taken as "brute" basic laws of the world—"brute" in the sense that they are not further explainable and must be taken to be among the fundamental laws of our total theory of the world. (We will shortly discuss an argument, "explanatory argument I," that claims that these psychoneural correlations are explained by psychoneural identities; for example, that "pain occurs iff Cfs occurs" is explained by "pain = Cfs.")

But such a theory of the world should strike us as intolerably complex and bloated—the very antithesis of simplicity and elegance we strive for in science. For one thing, it includes a huge and motley crowd of psychoneural correlation laws—a potentially infinite number of them—among its basic laws. For another, each of these psychoneural laws is highly complex: Pain may be a "simple" sensory quality, but look at the physical side of the pain-Cfs correlation. Cfs consists of an untold number of molecules, atoms, and particles, and their interactions. We expect our basic laws to be reasonably simple, and reasonably few in number. And we expect to explain complex phenomena by combining and iteratively applying a few simple laws. We do not expect basic laws to deal in physical structures consisting of zillions of particles in unimaginably complex configurations. This makes our total theory messy, inflated, and inelegant.

Compare this bloated picture with what we get if we move from psychoneural correlations to psychoneural identities—from "pain occurs iff Cfs occurs" to "pain = Cfs." Pain and Cfs are one and not two, and we are not faced by two distinct phenomena whose correlation needs to be explained. In this way, psychoneural identities permit us to *transcend* and *renounce* these would-be correlation laws—what Herbert Feigl aptly called "nomological danglers."[7] Moreover, as Smart emphasizes, the identification of the mental with the physical brings the mental within the purview of physical theory, and ultimately our basic physics constitutes a complete and comprehensive explanatory framework adequate for all aspects of the natural world. The resulting picture is far simpler and more elegant than the earlier picture in which any complete theory of the world must include all those complex mind-brain laws in addition to the basic laws of physics. Anyway, that is the argument.

What should we think of this argument? Does going from psychoneural correlations to psychoneural identities really simplify our total theory of the world, as the argument claims? Here the reader is invited to reflect on the following simple question: Doesn't the psychoneural identity theory *merely*

replace psychoneural correlations with an *equal* number of psychoneural identities, one for one? The identities are empirical just like the correlations, and they make even stronger modal assertions about the world, going beyond the correlations. This is so because the identity "pain = Cfs" is now generally taken to be a necessary truth (if true), and the correlation "pain occurs iff Cfs occurs," being entailed by a necessary truth, turns out itself to be a necessary truth. Moreover, these identities are not deducible from more basic physical-biological laws any more than the correlations are, and so they must be countenanced as fundamental and ineliminable postulates about how things are in the world. So don't we end up with the same number of empirical assumptions about the world? The fact is that the total empirical content of a theory with psychoneural identities is at least equal to that of a theory with the psychoneural correlations they replace. Doesn't it follow that version (II) of the simplicity principle actually argues *against* psychoneural identities, or declares a tie between the identities and the correlations? So what exactly are the vaunted benefits of simplification promised by the identities?

The reader is also invited to consider how a Cartesian, or a dualist of any stripe, might respond to Smart's simplicity argument, keeping in mind that one person's "simple" theory may well be another person's "incomplete" or "truncated" theory. What counts as "going beyond necessity" can be a matter of dispute—in fact, what is to be included among "the necessities" is usually the very bone of contention between the disputants.

Explanatory Arguments for Psychoneural Identity

According to some philosophers, psychoneural identities can do important and indispensable explanatory work—that is, they help explain certain facts and phenomena that would otherwise remain unexplained, and this provides us with a sufficient warrant for their acceptance. Sometimes an appeal is made to the principle of "inference to the best explanation." This principle is usually taken as an inductive rule of inference, and there is a widespread, if not universal, agreement that it is an important rule used in the sciences to evaluate the merits of theories and hypotheses. The rule can be stated something like this:

> *Principle of Inference to the Best Explanation*. If hypothesis H gives the *best* explanation of phenomena in a given domain when compared with other rival hypotheses $H_1, \ldots, H_n$, we may accept H as true, or at least we should prefer H over $H_1, \ldots, H_n$.[8]

It is then argued that psychoneural identities, like "pain = Cfs," give the best explanations of certain facts, better than the explanations afforded by rival theories. The conclusion would then follow that the mind-body identity theory is the preferred perspective on the mind-body problem.

This argument comes in two versions, which diverge from each other in several significant ways. We consider them in turn.

Explanatory Argument I

The two explanatory arguments differ on the question of what it is that is supposed to be explained by psychoneural identities—that is, on the question of the "explanandum." Explanatory argument I takes the explanandum to be psychoneural correlations, claiming that psychoneural identities give the best explanation of psychoneural correlations. As we will see, explanatory argument II claims that the identities, rather than explaining the correlations, explain certain other facts about mental phenomena that would otherwise go unexplained. Let us see how the first explanatory argument is supposed to work.

First, it is claimed that specific psychoneural identities, like "pain = Cfs" and "consciousness = pyramidal cell activity," explain the corresponding correlations, like "pain occurs iff Cfs occurs" and "a person is conscious iff pyramidal cell activity is going on in the brain." As an analogy, consider this: Someone might be curious why Clark Kent turns up whenever and wherever Superman turns up. What better, or simpler, explanation could there be than the identity "Clark Kent *is* Superman"?[9] So the proponents of this form of explanatory argument claim that the following is an explanation of a psychoneural correlation and that it is the best available explanation of it:

(α) Pain = Cfs.
Therefore, pain occurs iff Cfs occurs.

Similarly for other psychological properties and their correlated neural properties.

Second, it is also claimed that the psychoneural identity theory offers the best explanation of the pervasive fact of psychoneural correlations, like this:

(β) For every mental property M there is a physical property P such that M = P.
Therefore, for every mental property M there is a physical property P such that M occurs iff P occurs.[10]

If we could show that psychoneural identities are the best explanations of psychoneural correlations, the principle of inference to the best explanation would sanction the conclusion that we are justified in taking psychoneural identities to be true, and that the psychoneural identity theory is the preferred position on the mind-body problem. Anyway, that is the idea.

But does the argument work? Obviously, specific explanations like (α) are crucial; if they do not work as explanations, there is no chance that (β), the explanation of the general mind-correlation thesis, will work. So is (α) an explanation? And is it the best possible explanation of the correlation? A detailed discussion of the second question would be a lengthy and time-consuming business: We would have to compare (α) with the explanations offered by epiphenomenalism, the double-aspect theory, the causal theory, and so on. But we can say this much in behalf of (α): It is ontologically the simplest. The reason is that all these other theories are dualist theories, and in consequence they have to countenance more entities—mental events in addition to brain events. But is (α) overall the *best* explanation? Fortunately, we can set aside this question because there are serious reasons to be skeptical about its being an explanation at all. If it is not an explanation, the question of whether it is the best explanation does not arise.

First consider this: If pain indeed is identical with Cfs, in what sense do they "correlate" with each other? For there is here only one thing, whether you call it "pain" or "Cfs," and as Smart says in the paragraph quoted earlier, you cannot correlate something with itself. For Smart, the very point of moving to the identity "pain = Cfs" is to transcend and cancel the correlation "pain occurs iff Cfs occurs." This is the "nomological dangler" to be eliminated. For it seduces us to ask wrongheaded and unanswerable questions like "Why does pain correlate with Cfs?" "Why doesn't itch correlate with Cfs?" "Why does any conscious experience correlate with Cfs?" and so on. By opting for the identity, we show that these questions have no answers, since the *presupposition* of the questions—namely, that pain *correlates* with Cfs—is false. The question "Why is it the case that *p*?" presupposes that *p* is true. When *p* is false, the question has no correct answer and it cancels itself as an explanandum. Showing that a demand for an explanation rests on a false presupposition is one way to deal with it; providing an explanation is not the only way.

A defender of the explanatory argument might protest our talk of "correlations," objecting that we are assuming, with Smart, that a "correlation" requires two distinct items. We should stop calling "pain occurs iff Cfs occurs" a *correlation*, if that is going to lead anyone to infer pain and Cfs to be two things. It is pointless to be hung up on the word "correlation." Whatever you call it, the fact

expressed by "pain occurs iff Cfs occurs" is explained by the identity "pain = Cfs," and, moreover, this is the best possible explanation of it. That is all we need to make the explanatory argument work.

It is doubtful, however, that this reply will get the explanatory argument out of trouble. In the first place, this move will not make questions like "Why does pain, not itch, correlate with Cfs?" go away. For we can readily reformulate it as follows: Why is it the case that pain occurs iff Cfs occurs, rather than itches occurring just when Cfs occurs? Would we take the following answer from the proponent of the explanatory argument as an acceptable explanation? "That's because pain is identical with Cfs but itch isn't identical with it." It is doubtful that most of us would consider this an informative answer—an informative explanation of why pains, but not itches, are associated with Cfs. Some notable thinkers, William James and T.H. Huxley among them, have long despaired of our ever being able to explain why these particular mind-body associations (or whatever you wish to call them) hold. The idea that simply by moving from mere associations to identities, we can resolve the explanatory puzzles of Huxley and James seems too good to be true.

Second, if it is true that pain = Cfs, the fact to be explained, namely that pain occurs iff Cfs occurs, is just the fact that pain occurs iff pain occurs, or that Cfs occurs iff Cfs occurs, and these manifestly trivial facts (if they are facts at all), with no content, seem neither in need of an explanation nor capable of receiving one. So rather than offering an explanation of why pain occurs just in case Cfs occurs, the proposal that pain = Cfs transforms the supposed explanandum into something for which explanation seems entirely irrelevant. Rather than explaining it, it disqualifies it as an explanandum.

As we have seen, the argument under consideration invokes the principle of inference to the best explanation as a scientific rule of induction; however, most explanations of correlations in the sciences seem to work quite differently. There appear to be two common ways of explaining correlations in science. First, scientists sometimes explain a correlation by deducing it from more fundamental correlations and laws (as when the correlation between the length and the period of swing of a simple pendulum is explained in terms of more basic laws of mechanics). Second, a correlation is often explained by showing that the two correlated phenomena are collateral effects of a common cause. (Recall the earlier example in which the correlation between the phases of the moon and tidal actions is explained in terms of the astronomical configurations involving the sun, the moon, and the earth; an explanation of co-occurrences of two medical symptoms on the basis of a single underlying disease.) It should be noticed that neither of these two ways renders the correlations into trivialities;

these explanations respect their status as correlations and provide serious and informative explanations for them. Indeed, it is difficult to think of a scientific example in which a correlation is explained by simply identifying the phenomena involved.

There is a further notable feature of scientific hypothesis testing: When a new hypothesis is proposed as the best explanation of the existing data, the scientists do not stop there; they will go on to subject the hypothesis to further tests, by deriving additional predictions and looking for new applications. When "pain = Cfs" is proposed as the best explanation of "pain occurs iff Cfs occurs," what *further* predictions can we derive from "pain = Cfs" for additional tests? Are there predictions, empirical or otherwise, derivable from this identity that are not derivable from the correlation "Cfs causes pain," or the emergent hypothesis "pain is an emergent phenomenon arising from Cfs," or the epiphenomenalist hypothesis "Cfs causes pain"? It seems clear that genuine scientific uses of the inference to the best explanation principle bears little resemblance to its use in explanatory argument I for psychoneural identities. The principle of inference to the best explanation gains credibility from its use in scientific hypothesis testing. Using it to support what is an essentially philosophical claim, with no predictive implications of its own and hence no possibility of further tests, seems at best a misapplication of the principle; it can mislead us into thinking that the choice of a position on the mind-body problem is like a quotidian testing of rival scientific hypotheses. Even J. J. C. Smart, arguably the most optimistic and stalwart physicalist ever, had this to say:

> If the issue is between (say) a brain-process thesis and a heart thesis, or a liver thesis, or a kidney thesis, then the issue is a purely empirical one, the verdict is overwhelmingly in favor of the brain. . . . On the other hand, if the issue is between a brain-or-liver-or-kidney thesis (that is some form of materialism) on the one hand and epiphenomenalism on the other hand, then the issue is not an empirical one. For there is no conceivable experiment which could decide between materialism and epiphenomenalism.[11]

Further, the following consideration will reinforce our claim that the arguments against explanatory argument I has nothing to do with exploiting an informal connotation of the word "correlation." Let us ask: Exactly how does (α) work as an explanation? Explanation is most usefully thought of as derivation—a logical derivation, or proof, of the explanandum from the explanatory premises. So, then, how might the conclusion "pain occurs iff Cfs occurs" be derived from "pain = Cfs"? In formal logic, there is no rule of inference that says "From

'X = Y' infer 'X occurs iff Y occurs'"—for good reason, since a nonlogical term like "occur" is not part of formal logic. Instead, what we standardly find are the following two rules governing identity:

Axiom schema: X = X
Substitution rule: From " . . . X . . . " and "X = Y," infer " . . . Y . . . "

The first rule says that in a proof you can always write down as an axiom any sentence of the form "X = X," like "Socrates = Socrates" and "3 + 5 = 3 + 5." The second rule allows you to put "equals for equals." To put it another way, if X = Y and something is true of X, the same thing must be true of Y. This is the rule that is of the essence of identity. These two rules suffice to fix the logical properties of identity completely.

The following seems to be the simplest, and most natural, way of deriving "pain occurs iff Cfs occurs" from "pain = Cfs":

(γ) Pain = Cfs.
Pain occurs iff pain occurs.
Therefore, pain occurs iff Cfs occurs.

The first line is the premise, a psychoneural identity. The second line is a simple tautology of sentential logic, an instance of "*p* iff *p*," where *p* is any sentence you please, and we may write down a tautology anywhere in a derivation. The third line, the desired correlation, is derived by substituting "Cfs" for the second occurrence of "pain" in this tautology, in accordance with the substitution rule. As you see, the work that the identity "pain = Cfs" does is to enable us to *rewrite* the contentless tautology, "pain occurs iff pain occurs," by putting equals for equals. That is, the conclusion "pain occurs iff Cfs occurs," is a mere rewrite of "pain occurs iff pain occurs" and is equally contentless. As a mere rewrite rule in (γ), the identity "pain = Cfs" does no explanatory work, and hence cannot earn its warrant from the rule of inference to the best explanation.

If you think that calling the identity a "rewrite rule" is off the mark, trivializing its explanatory contributions, never mind what work the identity does in (γ); just consider this question: Does this derivation look to you like an explanation, a real explanation of anything? Now that you have (γ) in hand, would you say to yourself, "Now I finally understand why pain, not itch, occurs just in case my C-fibers are stimulated. I should tell my neuroscience professor about my discovery tomorrow!"? It seems as though once you recognize the pain-Cfs correlation as something to be explained, something you want to understand,

saying that they are one and the same thing will not meet your explanatory need. You will still wonder why pain, not itch, is identical with Cfs—which seems to take you back to the original question: Why does pain, not itch, co-occur with Cfs?

The role of identities in explanations is not well understood; there has been little informative discussion of this issue in the literature. Further, the view that explanation is fundamentally, or always, a derivational process is not universally accepted. However, the concept of explanation is deeply complex and difficult to pin down, and viewing explanatory processes as consisting in derivational activities is one of the few reasonably firm handles we have on this concept. If the defender of the explanatory argument insists that the explanation she has in mind of "pain occurs iff Cfs occurs" in terms of "pain = Cfs" does not proceed as a derivation, she is welcome to tell us exactly how she conceives of her explanation. That is, she needs to tell us just how the identity manages to explain its associated correlation.

There are reasons, then, to remain unpersuaded by the claim that psychoneural identities explain psychoneural correlations, and that for this reason the identities should be accepted as true.

Explanatory Argument II

This version of the explanatory argument does not claim that mind-body identities explain mind-body correlations; rather, they enable us to explain certain facts about mentality that would otherwise remain unexplained. How might we explain the fact that pain causes a feeling of distress? What is the causal mechanism involved? Suppose we have available the following psychoneural identities:

Pain = Cfs.
Distress = neural state N.

We might then be able to formulate the following neurophysiological explanation of why pain causes distress:

(θ) Neurophysiological laws
Cfs causes neural state N.
(I_1) Pain = Cfs.
(I_2) Distress = neural state N.
Therefore, pain causes distress.

Neurophysiological laws explain why Cfs causes N, and from this we derive our explanandum "Pain causes distress," by putting equals for equals on the basis of the psychoneural identities, (I_1) and (I_2). These identities help us explain a psychological regularity in terms of its underlying neural mechanism, and this seems just the kind of deeper scientific understanding we seek about higher-level psychological regularities.

Compare this with the situation in which we refuse to enhance correlations into identities. The best we could do with correlations would be something like this:

(λ) Neurophysiological laws
Cfs causes neural state N.
(C_1) Pain occurs iff Cfs occurs.
(C_2) Distress occurs iff neural state N occurs.
Therefore, pain correlates with a phenomenon that causes a phenomenon with which distress correlates.

This is no explanation of why pain causes distress; it doesn't even come close. To explain it, we need identities (I_1) and (I_2); correlations (C_1) and (C_2) will not do. According to the friends of this form of the explanatory arguments, an explanatory role of the kind played by psychoneural identities, as in (θ), yields sufficient justification for their acceptance.

Ned Block and Robert Stalnaker, proponents of the explanatory argument of this form, agree with J. J. C. Smart in regarding identities not as explaining their associated correlations but as helping us to get rid of them. They put the point this way:

> If we believe that heat is correlated with but not identical to molecular kinetic energy, we should regard as legitimate the question why the correlation exists and what its mechanism is. But once we realize that heat is molecular kinetic energy, questions like this will be seen as wrongheaded.[12]

Similarly, for "pain occurs iff Cfs occurs" and "pain = Cfs." The identity helps us avoid the "wrongheaded" question "Why does pain correlate with Cfs, not with something else?" by ridding us of the correlation. It is clear that contrary to the claims of explanatory argument I, Block and Stalnaker do not believe that this improper question is answered by the identity "pain = Cfs." We may summarize Block and Stalnaker's argument in favor of psychoneural identities as

follows: These identities *enable* desirable psychological explanations while *disabling* the improper demands for explanation of psychoneural correlations.[13]

How good is this argument? Unfortunately, not very good: The argument turns out to be problematic, for reasons similar to those that made explanatory argument I questionable. The trouble is that in both arguments the identities in question do not seem to do any explanatory work and hence are not qualified to benefit from the principle of inference to the best explanation. We can accept the claim that derivation (θ) gives a neurophysiological explanation of why pain causes distress: Laws of neurophysiology directly explain why Cfs causes neural state N, and given the identities "pain = Cfs" and "distress = neural state N," we would be justified in claiming that neurophysiological laws explain the fact that pain causes distress. This is so because, given the two identities, the statements "pain causes distress" and "Cfs causes neural state N" state one and the same fact. There is here one fact described in two ways—in the vernacular vocabulary and in the scientific vocabulary.

This shows just what goes wrong with explanatory argument II: The identities "pain = Cfs" and "distress = neural state N" do *no explanatory* work in this derivation. Their role is to enable us to *redescribe* a fact that has already been explained. The explanatory activity is over and finished at the second line when "Cfs causes neural state N" has been derived from, and thereby explained by, laws of neurophysiology. What the identities do is allow us to *rewrite* "Cfs causes neural state N" as "pain causes distress," by putting equals for equals. This is useful in presenting our explanatory accomplishment in neuroscience in the familiar "folk" language, but this involves no *explanatory* activity. The verdict, therefore, seems inescapable: Since the psychoneural identities have no involvement in explanation, they are ineligible as beneficiaries of the principle of inference to the best explanation. If there is a beneficiary of this principle in this situation, it is the laws of neuroscience because they do the explanatory work!

Our conclusion, therefore, has to be that both forms of the explanatory argument are vulnerable to serious objections. Their shared weakness is a lack of clear appreciation of just what role the psychoneural identities play in the explanations in which they supposedly figure. Our main contention has been that both arguments invoke, but misapply, the rule of inference to the best explanation, a principle that itself is far from uncontroversial.

An Argument from Mental Causation

By mental causation we mean any causal relation involving a mental event. A pin is run into your palm, causing you a sharp pain. The sudden pain causes

you to cry out and quickly pull back your hand. It also causes a feeling of distress and a desire to be rid of it. Causal relations involving mental and physical events are familiar facts of our everyday experience.

But pains do not occur without a physical basis; let us assume that pains are lawfully correlated with neural state N. So the sharp pain that caused the withdrawal of your hand has an occurrence of N as its neural substrate. Is there any reason for not regarding the latter, a neural event, as a cause of your hand's jerky motion?

Suppose we try to trace the causal chain backward from your hand's movement. The jerky motion was presumably caused by the contraction of muscles in your arm, which in turn was caused by neural signals reaching the muscles. The movement of neural signals is a complex physical process involving electrochemical interactions, and if we keep tracing the series of events backward to its source, we can expect it to culminate in a region in the central nervous system, perhaps in the cortex. Now ask yourself: Will this chain ever reach, or go through, a mental experience of pain, the pain you experienced when the pin was stuck in your palm? What could the transition from a neural event to a nonphysical, private pain event be like? Or the transition from a private pain experience to a public physicochemical neural event? How can a pain experience affect the motion of even a single molecule—speeding it up or slowing it down, or changing its direction? How can that happen? Is it even conceivable? It boggles our imagination!

The chances are that the causal chain culminating in your hand's jerky movement, when traced backward, will completely bypass your pain; there will be more and more neural-physical events as you keep going back, but no mental experiences. Nor does it make sense to postulate a purely mental causal chain, independent of the neural-physical chain, somehow reaching your muscles. (That's known as telekinesis—an alleged "psychic" phenomenon involving a mind causing a physical change at a distance, like bending a spoon by intensely gazing at it.) It seems, then, that the only way to salvage the pain as a cause of your hand motion is to think of it as a neural event. Which neural event? The best and most natural choice is its neural substrate, N (as we supposed), the state that is necessary and sufficient for the occurrence of the pain. This in brief is the causal argument, somewhat informally presented, for identifying mental states, especially states of consciousness, with neural states.

There is a more systematic, and currently influential, version of the causal argument that will now be presented. It begins with a premise asserting that mental causation is real:

i. Mental phenomena have effects in the physical world.

In this context, we take (i) as uncontroversial. Our beliefs and desires surely have the power to move our limbs and thereby enable us to cause things around us to be rearranged—moving the books from my desk to the bookshelves, emptying a waste basket, digging my car out of a snowbank, and starting an avalanche. If our mental states had no causal powers to affect physical things and events around us, we would cease to be agents, only helpless spectators of the passing scene. If that were true, our self-conception of ourselves as effective agents in the world would suffer a complete collapse.

Here is the second premise:

ii. [The causal closure of the physical domain] The physical world is causally closed. That is, if any physical event is caused, it has a sufficient physical cause (and a wholly physical causal explanation).

According to this principle, the physical world is causally self-contained and self-sufficient. It doesn't say that every physical cause has a sufficient physical cause—that is the principle of physical causal determinism. So (ii) is compatible with indeterminism about physical events. What (ii) says is that for any physical event, if we were to trace its causal ancestry, this need never take us outside the physical world. If a physical event has no physical cause, then it has no cause at all and no causal explanation. Further, this principle is compatible with dualism and other forms of nonphysicalism: As far as it goes, there could be a Cartesian world of immaterial minds, alongside the physical world, and all sorts of causal relations could hold in that world. The only thing, according to physical causal closure, is that the physical world must be causally insulated from such worlds; there can be no injection of causal influence into the physical world from outside. This means that there can be no "miracles" brought about by some transcendental, supernatural causal agents from outside physical space-time.

On Descartes's interactionist dualism, the physical causal closure fails: When an immaterial soul makes the pineal gland vibrate, thereby setting in motion a chain of bodily events, the motion of the pineal gland is caused, but it has no physical cause and no physical explanation. And this means that our physical theory would remain forever incomplete in the sense that there are physical events whose occurrences cannot be physically explained. A complete theory of the physical world would require references to nonphysical, immaterial causal agents and forces.

Why should we accept the causal closure of the physical domain? We will enumerate some reasons here without going into great detail.[14] First, there is

the widely noted success of modern science, in particular theoretical physics, which we take to be our basic science. Physics is all-encompassing: Nothing in the space-time world falls outside its domain. If a physicist encounters a physical event for which there is no ready physical explanation, or physical cause, she would consider that as indicating a need for further research; perhaps there are as-yet undiscovered physical forces. At no point would she consider the possibility that some nonphysical force outside the space-time world was the cause of this unexplained physical occurrence. The same seems to be true of research in other areas of science—broadly physical science including chemistry, biology, geology, and the like. If a brain scientist finds a neural event that is not explainable by currently known facts in neural science, what is the chance that she would say to herself, "Maybe this is a case of a Cartesian immaterial mind interfering with neural processes, messing up my experiment. I should look into that possibility!" We can be sure that would never happen. What would such research, investigating the workings of immaterial souls, look like? Where would you start? It isn't just that the principle of physical causal closure is the operative assumption in scientific research—remember that in science success is what counts. It may well be that there is a conceptual incoherence in the idea that there are nonphysical causal forces outside space-time that can causally intervene in what goes on in the space-time world.[15]

From these two premises, (i) and (ii), we have the desired conclusion:

(i) Mental phenomena are physical phenomena.

You might point out, rightly, that the only proposition we are entitled to derive is that only those mental phenomena that cause physical events are physical events.[16] Strictly speaking, that is correct, but remember this: Causation is transitive—that is, if one event causes another, and this second event causes a third, then the first event causes the third. If a mental event causes another mental event, which causes a physical event, the first event causes this physical event, and our argument pronounces it to be a physical event. Such chains of mental events can be as long as you wish; as long as a single event in this chain causes a physical event, every event preceding it in the chain qualifies as a physical event. This should pretty much cover all mental events; it is hard to imagine a mental causal chain consisting exclusively of mental events not touching anything physical anywhere. Even if there were such exceptions, the main physicalist point is made. A qualified conclusion stands: Mental events that have effects in the physical domain are physical events. The pain that causes your hand to pull back in a jerky motion and makes you cry

"Ouch!" is a physical event. But which physical event? What better candidate is there than the brain state that is the neural correlate of pain, namely Cfs? Cfs is a necessary and sufficient condition for the occurrence of pain, and it occurs exactly at the same time as the pain.

If in spite of these considerations you still want to insist on the pain as a separate cause of the hand movement, think of a new predicament in which you will find yourself. For the hand movement would now appear to have two distinct causes, the pain and its neural correlate Cfs, each presumably sufficient to bring it about. Doesn't that make this (and every other case of mental-to-physical causation) a case of causal overdetermination, an instance in which two independent causes bring about a single effect? Given that the hand withdrawal has a sufficient physical cause, namely Cfs, what *further* causal contribution can the pain make? There seems no leftover causal work that the pain has to be called on to perform. Again, the identification of the pain with Cfs appears to dissolve all these puzzles. There is, of course, the epiphenomenalist solution: Both the hand withdrawal and the pain are caused by Cfs, and the pain itself has no further causal role in this situation. But unlike the identity solution, the epiphenomenalist move renders the pain causally inert and ends up rejecting our initial assumption that a sharp pain caused the hand's jerky motion.

Perhaps a reconsideration of that assumption may be in order. The identification of a conscious pain experience with some molecular physical processes in the brain strikes some people as totally incredible and still others as verging on incoherence. If given a choice between taking pain and other experiences as physical processes in the brain on one hand and their causal impotence on the other, some may well consider the latter a preferable option. At this point, what the causal argument does is to give us a choice between psychoneural identity and epiphenomenalism: If you want to protect mental events from epiphenomenalism, you had better identify them with physical processes in the brain. To some, this may seem tantamount to discarding what is distinctively mental in favor of molecular physical processes in the body. On the other hand, if you are unwilling to embrace psychophysical identity, you put the causal powers of mentality in jeopardy. What good is our mentality if it is epiphenomenal? We will return to some of these issues later (chapter 7).

Against Psychoneural Identity Theory

There are three main arguments against the mind-brain identity theory. They are the epistemological argument, the modal argument, and the multiple realization argument. We consider each in turn.

The Epistemological Argument

Epistemological Objection 1. There is a group of objections based on the thought that the mental and the physical differ in their epistemological properties. Let us begin with the simplest, and rather simplistic, one. Medieval peasants knew lots about pains but nothing about C-fibers, and in fact little about the brain. So how can pains be identical with C-fiber excitations?

This objection assumes that the two statements "S knows something about X" and "X = Y" together entail "S knows something about Y." But is this true? It appears false: The same peasants knew a lot about water but nothing about H_2O. But that doesn't make the identity "water = H_2O" false. Suppose the objector persists: The peasants did know something about H_2O; after all, they knew a lot about water, and water *is* H_2O! How should we respond? Perhaps there is a sense in which the medieval peasants knew something about H_2O—we can concede that—but this must be a sense of knowing in which it is possible to know something about X without having the concept of X, or the ability to use the concept in forming thoughts or making judgments, or to use the expression "X" to express beliefs. But in this pale sense of knowing, there would be nothing wrong about saying that the peasants knew something about C-fiber excitation. They knew about C-fiber excitation in the same harmless sense in which they knew about H_2O. So the objection fails.

Epistemological Objection 2. According to the identity theory, specific psychoneural identities (for example, "pains are C-fiber excitations") are empirical truths discovered through scientific observation and theoretical research. If "$D_1 = D_2$" is an empirical truth, the two names or descriptions, D_1 and D_2, must have *independent criteria of application*. Otherwise, the identity would be a priori knowable; consider, for example, identities like "bachelor = unmarried adult male" and "the husband of Xanthippe = Xanthippe's male spouse." When an experience is picked out by a subject as a pain rather than an itch or tingle, the subject must do so by *recognizing*, or *noticing*, a certain distinctive felt character, a "phenomenal" or experiential quality, of the occurrence—its painful, hurtful quality. If pains were picked out by neurophysiological criteria (say, if we used C-fiber excitation as the criterion of pain), the identity of pain with a neural state could not be empirical; it would simply follow from the very criterion governing the concept of pain. This means, the objection goes, that to make sense of the supposed *empirical* character of psychoneural identities, we must acknowledge the existence of phenomenal, qualitative characters of experience distinct from neural properties.[17]

It seems, therefore, that the psychoneural identity physicalist still has these qualitative, phenomenal features of experience to contend with; even to make sense of her theory, there must be these nonphysical, qualitative properties by which we identify conscious experiences. It seems that she must somehow show that subjects do not identify mental states by noticing their qualitative features. Could the type physicalist argue that although a person does identify her experience by noticing its qualitative phenomenal features, they are not irreducible, since phenomenal properties as mental properties are identical, on her view, with physical-biological properties? But this reply is not likely to satisfy many people; it will invite the following response: "But surely when we notice our pains as pains, we do not do that by noticing biological or neural features of our brain states!" We immediately distinguish pains from itches and tickles; if we identified our experiences by their neurophysiological features, we should be able to tell which neurophysiological features represent pain, which represent itches, and so on. But is this credible?

Some philosophers have tried to respond to this question by analyzing away phenomenal properties. For example, Smart attempts to give phenomenal properties "topic-neutral translation."[18] According to him, when we say, "Adam is experiencing an orangish-yellow afterimage," the content of our report may be conveyed by the following "topic-neutral" translation—topic-neutral because it says nothing about whether what is being reported is mental or physical:

> Something is going on in Adam that is like what goes on when he is looking at an orangish-yellow color patch illuminated in good light.

(We suppose "looking" is explained physically in terms of his being awake, his eyes' being open and focused on the color patch, and so on.) Smart would add that this "something" that is going on in Adam is a brain state.

But will this satisfy someone concerned with the problem of explaining how someone manages to identify the kind of experience she is having? There is perhaps something to be said for these translations if we approach the matter strictly from the third-person point of view. But when you are reporting your own experience by saying, "I have a sharp pain in my left thumb," are you saying something like what Smart says that you are? To know that you are having an orangish-yellow afterimage, do you need to know anything about what generally goes on whenever you look at orangish-yellow color patches?

A more recent strategy that has become popular with latter-day-type physicalists is to press *concepts* into service and have them replace talk of properties

in the foregoing objection. The main idea is to concede *conceptual* differences between the mental and the neural but deny that these differences point to ontological differences, that is, differences in the properties to which these concepts apply or refer. This way of attempting to meet the objection is called the "phenomenal concept strategy." When we say that a person notices a pain by noticing its painfulness, this does not mean that the pain has the *property* of painfulness; rather, it means that she is "conceptualizing" her experience under the phenomenal *concept* of being painful—but the experience so conceptualized remains a neural state. The phenomenal concept is not a neural or physical concept; in particular, it is not identical with the concept of C-fiber stimulation. There is no consensus on what phenomenal concepts are; some take them as a type of "recognitional concept," like the concept red, which we apply to things on the basis of direct acquaintance with them; others take them to be a kind of demonstrative concept, like "*this* kind of experience," demonstratively referring to an experience of pain; there are many other views.[19] The main point is that a single property, presumably a physical-neural property, is picked out by both a phenomenal and a neural concept. Thus, we have a dualism of concepts, mental and physical, but a monism of properties, the entities referred to by these concepts. The advantage of framing the issues in terms of phenomenal concepts rather than phenomenal properties is supposed to derive from the fact that properties, whether phenomenal or of other sorts, are "out there" in the world, whereas concepts are part of our linguistic-conceptual apparatus for representing and describing what is out there. The strategy, then, is to take the phenomenal-neural differences out of the domain of facts of the world and bring them into the linguistic-conceptual domain. This, at any rate, is a move that has been made by some physicalists and it is currently receiving much attention in the field. Whether it is an essentially verbal ploy or something that is more substantial remains to be seen.

Epistemological Objection 3. Your knowledge that you are thinking about an upcoming trip to East Asia is direct and private in the way that only first-person knowledge of one's own mental states can be. Others have to make inferences based on evidence and observation to find out what you are thinking, or even to find out that you are thinking. But your knowledge is not based on evidence or inference; somehow you directly know. In contrast, you have no such privileged access to your brain states. Your neurologist and neurosurgeon have much better knowledge of your brain than you do. In brief, mental states are directly accessible by the subject; brain states—and physical states in general—are not so accessible. So how can mental states be brain states?

We should note that for this objection to work, it is not necessary to claim that the subject has *infallible* access to all her mental states. For one thing, infallibility or absolute certainty is not the issue; rather, the issue is *private direct* access—that is, first-person access not based on inference from evidence or observation, the kind of access that no other person has. For another, it is only necessary that the subject have such access to at least *some* of her mental states. If that is the case, these mental states, according to this argument, cannot be identified with brain states, states for which public access is possible.

The identity theorist has to deny either the claim that we have direct private access to our own current mental states or the claim that we do not have such access to our brain states. She might say that when we know that we are in pain, we do have epistemic access to our Cfs, but our knowledge is under the description, or concept, "pain," not under the description "Cfs." Here there is one thing, Cfs (that is to say, pain), that can be known under two "modes of presentation"—pain and Cfs. Under one mode, the knowledge is private; under the other, it is public. It is like the same person is known both as "the husband of Xanthippe" and "the drinker of hemlock." You may know Socrates under one description but not the other. So knowledge is relative to the mode of description or conceptualization. Certain brain states, like Cfs, can be known in two different modes or under two different sorts of concepts, mental and physical. Knowledge under one mode can be different from knowledge under the other, and they need not co-occur. So this reply is in line with the final physicalist reply to epistemological objection 2, discussed earlier, which invoked phenomenal concepts. These replies, therefore, will likely stand or fall together.

In considering the viability of this reply, we can grant the point that knowledge and belief do depend on "modes of presentation" or ways of conceptualization or description. This seems like a plausible, and true, claim. What we ought to press for answers and elucidation is the following group of questions: Why is there a class of concepts or modes of presentation that gives rise to a very special type of knowledge, that is, knowledge by direct private access? There seems to be a philosophically important difference between such knowledge and our sundry knowledge of physical objects and events. What characteristics of this distinguished class of concepts and these modes of presentation explain the fact that they allow this special type of knowledge? If we conceptualize C-fiber stimulation under the mental concept "itch," that would presumably be wrong. Why? What makes it wrong? The dualist seems to have a simple perspective on these issues: These mental concepts and modes of presentation apply to, or signify, mental events that are directly and privately accessible to the subject; there is not, nor need there be, anything special about the concepts and

modes of presentation themselves. This is exactly the kind of reply that the psychoneural identity theorist wants to avoid.

The Modal Argument

Type physicalists used to say that mind-brain identities—for example, "pain = C-fiber activation"—are *contingent*, not necessary. That is, although pain is in fact C-fiber excitation, it could have been otherwise; there are possible worlds in which pain is not C-fiber excitation but some other brain state—perhaps not a brain state at all. The idea of contingent identity can be explained by an example such as this: "Barack Obama is the forty-fourth president of the United States." The identity is true, but it might have been false: There are possible worlds in which the identity does not hold—for example, one in which Obama decided to pursue an academic career rather than politics, one in which Senator Hillary Clinton won the Democratic nomination, one in which Senator John McCain defeated Obama, and so on. In all these worlds someone other than Barack Obama would be the forty-fourth president of the United States.

But this is possible only because the expression "the forty-fourth president of the United States" can refer to different persons in different possible worlds; things might have gone in such a way that the expression designated someone other than Obama—for example, Hillary Clinton or John McCain. Expressions like "the forty-fourth president of the United States," "the 2009 Wimbledon Men's Singles Champion," and "the tallest man in China," which can name different things in different possible worlds, are what Saul Kripke calls "nonrigid designators."[20] In contrast, proper names like "Barack Obama," "Socrates," and "Number 7" are "rigid"—they designate the same objects in all possible worlds in which they exist. The forty-fourth president of the United States might not have been the forty-fourth president of the United States (for example, if Obama had lost to Clinton), but it is not true that Barack Obama might not have been Barack Obama. (Obama might not have been called "Barack Obama," but that is another matter.) This shows that a contingent identity, "X = Y," is possible only if either of the two expressions, "X" or "Y," is a nonrigid designator, an expression that can refer to different things in different worlds.

Consider the term "C-fiber excitation": Could this designator be nonrigid? It would seem not: How could an event that in fact is the excitation of C-fibers not have been one? How could an event that is an instance of C-fiber excitation be, say, a volcano eruption or a collision of two stars in another possible world? A world in which no C-fiber excitation ever occurs is a world in which this event, which is a C-fiber excitation, does not occur. The term "pain" also

seems rigid. If you are inclined to take the painfulness of pain as its essential defining property, you will say that "pain" rigidly designates an event or state with this quality of painfulness and that the expression designates an event of that sort across all possible worlds. A world in which nothing ever hurts is a world without pain.

It follows that if pain = Cfs, then this must be a necessary truth—that is, it must hold in every possible world. Descartes famously claimed that it is possible for him to exist as a thinking and conscious thing even without a body. If that is possible, then pain could exist even if Cfs did not. Some philosophers have argued that "zombies"—creatures that are physically just like us but have no consciousness—are possible; that is, there are possible worlds inhabited by zombies. If so, Cfs could exist without being accompanied by pain. If these are real possibilities, then "pain = Cfs" cannot be a necessary truth. Then, by the principle that if X and Y are rigid designators, the identity "X = Y" is necessarily true, if true, it follows that "pain = Cfs" is false—false in this world. More generally, psychoneural identities are all false.

Many mind-brain identity theorists would be likely to dispute the claim that it is possible that pain can exist even if Cfs does not, and they would question the claim that zombies are a real possibility. We can grant, they will argue, that in some sense these situations are "conceivable," that we can "imagine" such possibilities. But the fact that a situation is conceivable or imaginable does not entail that it is genuinely possible. For example, it is conceivable, they will say, that water is not H_2O and that heat is not molecular kinetic energy; the concept of water and the concept of H_2O are logically unrelated to each other, and there is no conceptual incoherence or contradiction in the thought that water ≠ H_2O. And we might even say that "water ≠ H_2O" is *epistemically possible*: For all that people knew about water and other things not so long ago, it was possible that water could have turned out to be something other than H_2O. That is, for all we knew a couple hundred years ago, we might be living on a planet with XYZ, rather than H_2O, coming out of the tap, filling our lakes and rivers, and so on, where XYZ is observationally indistinguishable from H_2O, although wholly different in molecular structure. Nonetheless, water = H_2O, and necessarily so. The gist of the reply by the identity theorists then is that conceivability does not entail real metaphysical possibility and that this is shown by a posteriori necessary identities like "water = H_2O" and "heat = molecular kinetic energy." For them, psychoneural identities, "pain = Cfs" and the like, are necessary a posteriori truths just like these scientific identities. Issues about conceivability and possibility are highly complex and contentious, and they are being actively debated, without a consensus resolution in sight.[21]

The Multiple Realization Argument

The psychoneural identity theory says that pain is C-fiber excitation. But that implies that unless an organism has C-fibers, it cannot have pain. But aren't there pain-capable organisms, like reptiles and mollusks, with nervous systems very different from the human nervous system? Perhaps in these species the cells that work as nociceptive neurons—pain-receptor neurons—are not like human C-fibers at all; how can we be sure that all pain-capable animals have C-fibers? Can the identity physicalist reply that it should be possible to come up with a more abstract and general physiological description of a brain state common to all organisms, across all species, that are in pain? This seems highly unlikely, and in any case, how about inorganic systems? Could there not be intelligent extraterrestrial creatures with a complex and rich mental life but whose biology is not carbon-based? And is it not conceivable—in fact, nomologically possible if not practically feasible—to build intelligent electromechanical robots to which we would be willing to attribute various mental states (perceptual and cognitive states, if not sensations and emotions)? Moreover, the neural substrates of highly specific mental states (e.g., having the belief that winters are colder in New Hampshire than in Rhode Island) can differ from person to person and may change over time even in a single person through maturation, learning, and brain injuries. Does it make sense to think that some single neural state is shared by all persons who believe that cats are smarter than dogs, or that 7 + 5 = 12? Moreover, we should keep in mind that if pain is identical with some physical state, this must hold not only in actual organisms and systems but in all possible organisms and systems. This is so because, as we saw earlier in our discussion of the modal argument, such identities, if true, must be necessarily true.

These considerations are widely thought to show that any mental state is "multiply realizable"[22] in a large variety of physical-biological systems, with the consequence that it is not possible to identify mental states with physical states. If pain is identical with a physical state, it must be identical with some *particular* physical state, but there is no single neural correlate or substrate of pain. On the contrary, there must be indefinitely many physical states that can "realize" (or "instantiate," or "implement") pain in all sorts of pain-capable organisms and systems. So pain, as a type of mental state, cannot be identified with a neural state type or with any other physical state type.

This is the influential and widely known "multiple realization argument" that Hilary Putnam and others advanced in the late 1960s and early 1970s. It has had a critical impact on the way philosophy of mind has developed since

then. It was this argument, rather than any of the other difficulties, that brought about an unexpectedly early decline of psychoneural identity theory. What made the multiple realization argument distinctive, and different from other sundry objections, was that it brought with it a fresh and original conception of the mental, which offered an attractive alternative approach to the nature of mind. This is functionalism, still the reigning orthodoxy on the nature of mentality and the status of psychology. We turn to this influential view in the next two chapters.

Reductive and Nonreductive Physicalism

The psychoneural identity theory, or identity physicalism, is a form of reductive physicalism. It reductively identifies mental states with neural states of the brain. It is also called type physicalism, since it identifies types, or kinds, of mental states, like pain, thirst, anger, and so on, with types and kinds of neural-physical states. That is, psychological types, or properties, are claimed to be identical with neural-physical types and properties. Thus, type physicalism contrasts with the so-called token physicalism (see chapter 1), according to which, though psychological types and properties are not neural-physical types, each individual, "token" psychological event, like this particular pain I am experiencing now, is in fact a neural event. This means that different tokens, or instances, of a single mental kind may, and usually will, fall under distinct neural kinds. Both you and an octopus experience a pain, but your pain is an instance of C-fiber stimulation and the octopus pain is an instance of (let's say) O-fiber stimulation. As you can tell, token physicalism is inspired by considerations of multiple realization of psychological states.

Since the 1970s, chiefly on account of the influence of the multiple realization argument, reductive physicalism has had a rough time of it, although of late it has shown renewed strength and signs of a revival. As reductionism's fortunes declined, nonreductive physicalism (see chapter 1) rapidly gained strength and influence, and it has reigned as the dominant and virtually unchallenged position on the mind-body problem for the past several decades. This is the view that mental properties, along with other "higher-level" properties of the special sciences, like biology, geology, and the social sciences, resist reduction to the basic physical domain. An antireductionist view of this kind has also served as an influential philosophical foundation of psychology and cognitive science, providing support for the claim that these sciences are autonomous, each with its own distinctive methodology and system of concepts and not answerable to the

methodological or explanatory constraints of more fundamental sciences. Thus, the most widely accepted form of physicalism today combines substance physicalism with property dualism: All concrete individual things in this world are physical, but complex physical systems can, and sometimes do, exhibit properties that are not reducible to "lower-level" physical properties. Among these irreducible properties are, most notably, psychological properties, including those investigated in the psychological and cognitive sciences.

But nonreductive physicalism, above all, is a form of physicalism. What makes it physicalistic? In what do its credentials as physicalism consist? Part of the answer is that it accepts substance physicalism. It rejects Cartesian mental substances and other supposed nonphysical things in space-time, and of course there is nothing outside space-time. Although the nonreductive physicalist denies the physical reducibility of the mental, she nonetheless accepts a close and intimate relationship between mental properties and physical properties, and this is mind-body supervenience (see chapter 1). We may call this *supervenience physicalism*. Some nonreductive physicalists will go a step further and maintain that their irreducible mental properties are "physically realized" or "physically implemented." This is the so-called *realization physicalism*.[23] There will be more in the next chapter on the idea of physical realization; here, the point to note is that the realization relation is stronger than supervenience, and hence that realization physicalism is a stronger thesis than supervenience physicalism. If mind-body realization holds, then mind-body supervenience holds, but not the other way around.

In any case, in committing herself to the supervenience, or realization, relation between mental and physical properties, the nonreductive physicalist goes beyond mere property dualism. It should be clear that property dualism as such does not require the thesis that the mental character of a being is dependent on, or determined by, its physical nature, as mind-body supervenience requires, or that mental properties are physically realized (if they are realized at all). Mental properties, though instantiated in physical systems, might yet be independent of their physical properties.

Moreover, nonreductive physicalists are mental realists who believe in the reality of mental properties; they regard mental properties as genuine properties the possession of which makes a difference—a causal difference. Part of the belief in the reality of mental properties is to believe in their causal efficacy. An organism, in virtue of having a mental property (say, wanting a drink of water or being in pain), acquires powers and propensities to act or be acted upon in certain ways. Summarizing all this, nonreductive physicalism, as standardly understood, comprises the following four claims:

Substance Physicalism. The space-time world consists exclusively of bits of matter and their aggregates.

Irreducibility of the Mental. Mental properties are not reducible to physical properties.

Mind-Body Supervenience or *Realization*. Either (a) mental properties supervene on physical properties, or (b) mental properties, when they are realized, are realized by physical properties.

Mental Causal Efficacy. Mental properties are causally efficacious; mental events are sometimes causes of other events, both physical and mental.

Nonreductive physicalism, understood as the conjunction of these four theses, has been the most influential position on the mental-physical relation. We can think of property dualism as the conjunction of the first, second, and fourth doctrines—that is, all but mind-body supervenience/realization. Besides its acceptance of substance physicalism, what makes nonreductive physicalism a serious physicalism is its commitment to mind-body supervenience/realization. Property dualism that rejects mind-body supervenience/realization seems, prima facie, to be a possible position; however, this form of property dualism has not found strong advocates and remains largely undeveloped. And there may be a good reason for this: In rejecting supervenience/realization, you take the mental as constituting its own realm separate from the physical and it is difficult to see how you would be able to explain the causal efficacy of the mental in the physical world. You might very well run into troubles of the kind Descartes had in explaining how immaterial minds could causally interact with material things (see chapter 2). The rejection of mind-body supervenience, therefore, may force you to give up mental causal efficacy, and this is not an option for most (see chapter 7).

In accepting the irreducibility thesis, nonreductive physicalism attempts to honor the special position that thought and consciousness enjoy in our conception of ourselves among the things of this world. As was noted above, the irreducibility thesis is also an affirmation of the autonomy of psychology and cognitive science as sciences in their own right, not constrained by more basic sciences. In accepting the causal efficacy of the mental, the nonreductive physicalist not only acknowledges what seems so familiar and obvious to common sense, but at the same time, it declares psychology and cognitive science to be

genuine sciences capable of generating law-based causal explanations and predictions. All in all, it is an attractive package, and it is not difficult to understand its appeal and staying power.

However, all that may only be wishful thinking. The story may be too good to be true. There have recently been significant objections and criticisms of the nonreductive aspect of nonreductive physicalism, and these collectively have generated enough pressure for many philosophers to reconsider its viability. We will see some of the difficulties nonreductive physicalism faces in regard to mental causation later (see chapter 7).

For Further Reading

The classic sources of the mind-brain identity theory are Herbert Feigl, "The 'Mental' and the 'Physical,'" and J. J. C. Smart, "Sensations and Brain Processes," both of which are available in *Philosophy of Mind: Classical and Contemporary Readings*, edited by David J. Chalmers. The Smart article is widely reprinted in anthologies on philosophy of mind. For more recent book-length treatments of physicalism and related issues, see Christopher S. Hill, *Sensations: A Defense of Type Physicalism*; Jeffrey Poland, *Physicalism: The Philosophical Foundation*; Andrew Melnyk, *A Physicalist Manifesto*; Thomas W. Polger, *Natural Minds*; Jaegwon Kim, *Physicalism, or Something Near Enough*; Daniel Stoljar, *Physicalism*.

For criticisms, see Saul Kripke, *Naming and Necessity*, lecture 3. John Heil's anthology, *Philosophy of Mind: A Guide and Anthology*, includes three essays (by John Foster, Peter Forrest, and E. J. Lowe) that are worth examining in a section with the title "Challenges to Contemporary Materialism." A very recent collection of critical essays on physicalism is *The Waning of Materialism*, edited by Robert C. Koons and George Bealer.

For the multiple realization argument against the psychoneural identity theory, the original sources are Hilary Putnam, "Psychological Predicates," later retitled "The Nature of Mental States," and Jerry Fodor's "Special Sciences, or the Disunity of Science as a Working Hypothesis." For recent reevaluations of the argument, see Jaegwon Kim, "Multiple Realization and the Metaphysics of Reduction"; William Bechtel and Jennifer Mundale, "Multiple Realizability Revisited: Linking Cognitive and Neural States." There is an extensive discussion of realization and multiple realizability in Lawrence Shapiro, *The Mind Incarnate*.

On the status of nonreductive physicalism, see Kim, "The Myth of Nonreductive Physicalism" and "Multiple Realization and the Metaphysics of Reduction"; Andrew Melnyk, "Can Physicalism Be Non-Reductive?" For responses:

Ned Block, "Anti-Reductionism Slaps Back"; Jerry Fodor, "Special Sciences: Still Autonomous After All These Years"; Louise Antony, "Everybody Has Got It: A Defense of Non-Reductive Materialism."

Notes

1. See René Descartes, *The Passions of the Soul*.

2. See Thomas H. Huxley, "On the Hypothesis That Animals Are Automata, and Its History."

3. Samuel Alexander, *Space, Time, and Deity*. Vol. 2, p. 47. "Natural piety" is an expression made famous by the poet William Wordsworth.

4. J. J. C. Smart, "Sensations and Brain Processes." U. T. Place's "Is Consciousness a Brain Process?" published in 1956, predates Smart's article as perhaps the first modern statement of the identity theory.

5. J. J. C. Smart, "Sensations and Brain Processes," p. 117 (in the reprint version in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil. Emphasis in the original).

6. See the entry "William of Ockham" in the *Macmillan Encyclopedia of Philosophy*, 2nd ed.

7. Herbert Feigl, "The 'Mental' and the 'Physical,'" p. 428.

8. See Gilbert Harman, "The Inference to the Best Explanation." For a critique of the principle, see Bas Van Fraassen, *Laws and Symmetry*.

9. This example comes from Christopher S. Hill, *Sensations: A Defense of Type Materialism*, p. 24. Hill's book includes an extremely clear and forceful presentation of explanatory argument I.

10. This is substantially the form in which Brian McLaughlin formulates his explanatory argument. See his "In Defense of New Wave Materialism: A Response to Horgan and Tienson." Hill (see note 10) and McLaughlin are two leading proponents of this form of the explanatory argument. However, McLaughlin does not explicitly invoke the rule of inference to the best explanation. See also Andrew Melnyk, *A Physicalist Manifesto*.

11. J. J. C. Smart, "Sensations and Brain Processes," p. 126.

12. Ned Block and Robert Stalnaker, "Conceptual Analysis, Dualism, and the Explanatory Gap," p. 24.

13. It is debatable whether it really is improper, or wrongheaded (as Block and Stalnaker put it), to ask for explanations of psychoneural correlations. One might argue that such explanatory demands are perfectly in order, and that to the extent that physicalism is unable to meet them, it is a limited and flawed doctrine.

14. For further discussion, see David Papineau, "The Rise of Physicalism" and *Thinking About Consciousness,* chapter 1.

15. You might recall the pairing problem discussed in chapter 2, in connection with Descartes's interactionist dualism.

16. There is another issue with the argument as presented, which is discussed in chapter 7 on mental causation; see the section on the "exclusion argument."

17. This objection is worked out in detail in Jerome Shaffer, "Mental Events and the Brain." The original form of this argument is credited to Max Black by J. J. C. Smart in his "Sensations and Brain Processes."

18. See J. J. C. Smart, "Sensations and Brain Processes."

19. This strategy originated in Brian Loar, "Phenomenal States." For more recent discussions, see *Phenomenal Concepts and Phenomenal Knowledge*, ed. Torin Alter and Sven Walter. Also helpful are: Katalin Balog, "Phenomenal Concepts," in *Oxford Handbook of Philosophy of Mind*, ed. Brian McLaughlin et al.; Peter Carruthers and Benedicte Veillet, "The Phenomenal Concept Strategy."

20. This neo-Cartesian modal argument is due to Saul Kripke. See his *Naming and Necessity*, especially lecture 3, in which the argument is presented in detail.

21. For further discussion of these issues, see the essays in *Conceivability and Possibility*, edited by Tamar Szabo Gendler and John Hawthorne.

22. The terms "variably realizable" and "variable realization" are commonly used by British writers.

23. I believe Andrew Melnyk first used this term in his *A Physicalist Manifesto*. Jaegwon Kim used the term "physical realizationism" earlier in *Mind in a Physical World*, but "realization physicalism" is better.

CHAPTER 5

Mind as a Computing Machine

Machine Functionalism

In 1967 Hilary Putnam published a paper of modest length titled "Psychological Predicates."[1] This paper changed the debate in philosophy of mind in a fundamental way, by doing three remarkable things: First, it quickly brought about the decline and fall of type physicalism, in particular, the psychoneural identity theory. Second, it ushered in functionalism, which has since been a highly influential—arguably the dominant—position on the nature of mind. Third, it was instrumental in installing antireductionism as the orthodoxy on the nature of psychological properties. Psychoneural identity physicalism, which had been promoted as the only view of the mind properly informed by the best contemporary science, turned out to be unexpectedly short-lived, and by the mid-1970s most philosophers had abandoned reductionist physicalism not only as a view about psychology but as a doctrine about all special sciences, sciences other than basic physics.[2] In a rapid shift of fortune, identity physicalism was gone in a matter of a few years, and functionalism was quickly enthroned as the "official" philosophy of the burgeoning cognitive science, a view of psychological and cognitive properties that best fit the projects and practices of the scientists.

All this stemmed from a single idea: *the multiple realizability of mental properties*. We have already discussed it as an argument against the psychoneural identity theory and, more generally, as a difficulty for type physicalism (chapter 4). What sets the multiple realization argument apart from numerous other objections to the psychoneural identity theory is that it gave birth to an

attractive new conception of the mental that has played a key role in shaping an influential view of the nature and status of not only cognitive science and psychology but also other special sciences.

Multiple Realizability and the Functional Conception of Mind

Perhaps not many of us now believe in angels—purely spiritual and immortal beings supposedly with a full mental life. Angels, as traditionally conceived, are wholly immaterial beings with knowledge and belief who can experience emotions and desires and are capable of performing actions. The idea of such a being may be a perfectly coherent one, like the idea of a unicorn or Bigfoot, but there seems no empirical evidence that there are beings fitting the description, just as there are no unicorns and probably no Bigfoot. So like unicorns but unlike married bachelors or four-sided triangles, there seems nothing conceptually impossible about angels. If the idea of an angel with beliefs, desires, and emotions is a consistent one, that would show that there is nothing in the idea of mentality as such that precludes purely nonphysical, wholly immaterial beings with psychological states.[3]

It seems, then, that we cannot set aside the possibility of immaterial realizations of mentality as a matter of an a priori conceptual fact.[4] Ruling out such a possibility requires commitment to a substantive metaphysical thesis, perhaps something like this:

> *Realization Physicalism*. If something *x* has some mental property M (or is in mental state M) at time *t*, then *x* is a physical thing and *x* has M at *t* in virtue of the fact that *x* has at *t* some physical property P that realizes M in *x* at *t*. [5]

It is useful to think of this principle as a way of stating the thesis of physicalism.[6] It says that anything that exhibits mentality must be a physical system—for example, a biological organism. Although the idea of nonphysical entities having mental properties may be a consistent one, the actual world is so constituted, according to this thesis, that only physical systems, like biological organisms, turn out to have mental properties—maybe because they are the only things that exist in space-time. Moreover, the principle requires that every mental property be physically based; each occurrence of a mental property is due to the occurrence of a physical "realizer" of the mental property. A simple way of putting the point would be this: Minds, if they exist, must be embodied.

Notice that this principle provides for the possibility of multiple realization of mental properties. Mental property M—say, being in pain—may be such that in humans C-fiber activation realizes it but in other species (say, octopuses and reptiles) physiological mechanisms that realize pain may be vastly different. Perhaps there might be non-carbon-based or non-protein-based biological organisms with mentality, and we cannot a priori preclude the possibility that electromechanical systems, like the "intelligent" robots and androids in science fiction, might be capable of having beliefs, desires, and even sensations. All this suggests an interesting feature of mental concepts: They seem to carry no constraint on the actual physical-biological mechanisms that realize or implement them. In this sense, psychological concepts are like concepts of artifacts. For example, the idea of an "engine" is silent on how an engine might be designed and built—whether it uses gasoline or electricity or steam and, if it is a gasoline engine, whether it is a piston or rotary engine, how many cylinders it has, whether it uses a carburetor or fuel injection, and so on. As long as a physical device is capable of performing a certain specified job—in this instance, that of transforming various forms of energy into mechanical force or motion—it counts as an engine. The concept of an engine is defined by a *job description*, or *causal role*, not a description of mechanisms that execute the job. Many biological concepts are similar: What makes an organ a heart is the fact that it pumps blood. The human heart may be physically very unlike hearts in, say, reptiles or birds, but they all count as hearts because of the job they do in the organisms in which they are found, not on account of their similarity in shape, size, or material composition.

What, then, is the job description of pain? The capacity for experiencing pain under appropriate conditions—in particular, when an organism suffers tissue damage—is critical to its chances for adaptation and survival. There are unfortunate people who congenitally lack the capacity to sense pain, and few of them survive into adulthood.[7] In the course of coping with the hazards presented by their environment, animal species must have had to develop pain mechanisms, "tissue-damage detectors," and it is plausible that different species, interacting with different environmental conditions and evolving independently, have developed different mechanisms for this purpose. As a start, then, we can think of pain as specified by the job description "tissue-damage detector"—a mechanism that is activated by tissue damage and whose activation in turn causes behavioral responses such as withdrawal, avoidance, and escape.

Thinking of the workings of the mind in analogy with the operations of a computing machine is commonplace, both in the popular press and in serious philosophy and cognitive science, and we will soon begin looking into the

mind-computer analogy in detail. A computational view of mentality also shows that we must expect mental states to be multiply realized. We know that any computational process can be implemented in a variety of physically diverse computing machines. Not only are there innumerable kinds of electronic digital computers (in addition to the semiconductor-based machines we are familiar with, think of the vacuum-tube computers of olden days), but also computers can be built with wheels and gears (as in Charles Babbage's original "Analytical Engine") or even with hydraulically operated systems of pipes and valves, although these would be unacceptably slow (not to say economically prohibitive). And all of these physically diverse computers can be performing "the same computation," say, solving the same differential equations. If minds are like computers and mental processes—in particular, cognitive processes—are, at bottom, computational processes, we should expect no prior constraint on just how minds and mental processes are physically implemented, that is, realized. Just as vastly different physical devices can execute the same computational program, so vastly different biological or physical systems should be able to subserve the same cognitive processes. Such is the core of the functionalist conception of the mind.

What these considerations point to, according to some, is the *abstractness* or *formality* of psychological properties in relation to physical or biological properties: Psychological kinds abstract from the physical and biological details of organisms so that states that are quite unlike from a physicochemical point of view can fall under the same psychological kind, and organisms and systems that are widely dissimilar biologically and physically can instantiate the same psychological regularities—or have "the same psychology." Psychological kinds seem to track *formal* patterns or structures of events and processes rather than their material constitutions or implementing physical mechanisms.[8] Conversely, the same physical structure, depending on the way it is causally embedded in a larger system, can subserve different psychological capacities and functions (just as the same computer chip can be used for different computational functions in the subsystems of a computer). After all, most neurons, it has been observed, are pretty much alike and largely interchangeable.[9]

What is it, then, that binds together all the physically diverse instances of a given mental kind? What do all pains—pains in humans, pains in canines, pains in octopuses, and pains in Martians—have in common in virtue of which they all fall under a single psychological kind, pain?[10] That is, what is the *principle of individuation* for mental kinds?

Let us first see how the type physicalist and the behaviorist answer this question. The psychoneural identity physicalist will say this: What all pains

have in common that makes them instances of pain is a certain neurobiological property, namely, being an instance of C-fiber excitation (or some such state). That is, for the type physicalist, a mental kind is a physical kind (a neurobiological kind, for the psychoneural identity theorist). You could guess how the behaviorist answers the question: What all pains have in common is a certain behavioral property—or to put it another way, two organisms are both in pain at a time just in case at that time they exhibit, or are disposed to exhibit, the behavior patterns characteristic of pain (for example, escape behavior, withdrawal behavior, and so on). For the behaviorist, then, a mental kind is a behavioral kind.

If you take the multiple realizability of mental states seriously, you will reject both these answers and opt for a "functionalist" conception. The main idea is that what is common to instances of a mental state must be sought at a higher level of abstraction. According to functionalism, a mental kind is a *functional kind*, or a *causal-functional kind*, since the "function" involved is to fill a certain causal role.[11] Let us go back to pain as a tissue-damage detector.[12] The concept of a tissue-damage detector is a *functional concept*, a concept specified by a job description, as we said: Any device is a tissue-damage detector for an organism just in case it can reliably respond to occurrences of damage to the tissues of the organism and transmit this information to other subsystems so that appropriate responses are produced. Functional concepts are ubiquitous: What makes something a mousetrap, a carburetor, or a thermometer is its ability to perform a certain function, not any specific physicochemical structure or mechanism; as someone said, anything is a mousetrap if it takes a live mouse as input and delivers a dead one as output. These concepts are specified by the functions that are to be performed, not by structural blueprints. As has been noted, many concepts, in ordinary discourse and in the sciences, are functional concepts in this sense; important concepts in chemistry and biology (for example, catalyst, gene, heart) seem best understood as functional concepts.

To return to pain as a tissue-damage detector: Ideally, every instance of tissue damage, and nothing else, should activate this mechanism and this must further trigger other mechanisms with which it is hooked up, leading finally to behavior that will in normal circumstances spatially separate the damaged part, or the whole organism, from the external cause of the damage. Thus, the concept of pain is defined in terms of its function, and the function involved is to serve as a *causal intermediary* between typical pain inputs (tissue damage, trauma, and so on) and typical pain outputs (winces, groans, avoidance behavior, and so on). Moreover, functionalism makes two significant additions. First, the causal conditions that activate the pain mechanism can include other

mental states (for example, you must be normally alert and not be absorbed in another activity, like intense competitive sports). Second, the outputs of the pain mechanism can include mental states as well (such as a sense of distress or a desire to be rid of the pain). Mental kinds are causal-functional kinds, and what all instances of a given mental kind have in common is that they all serve a certain *causal role* distinctive of that kind. And that is all. One might say that a functional kind has only a "nominal essence," given by its defining causal role, but no "real essence," a "deep" common property shared by all actual and possible instances of it.[13] Contrast this with water: All samples of water, anywhere anytime, must be quantities of H_2O molecules, and being composed of H_2O molecules is the essence of water. Pain does not have an essence in that sense. Functionalism itself may be characterized by the following slogan: "Psychological kinds have only nominal essences; they have no real essences."

In general, then, as David Armstrong has put it, the concept of a mental state is the concept of an internal state apt to be caused by certain sensory inputs and apt to cause certain behavioral outputs. A specification of input and output, <i, o>, will define a particular mental state: for example, <tissue damage, aversive behavior> defines pain, <skin irritation, scratching> defines itch, and so on.

Functional Properties and Their Realizers: Definitions

It will be useful to have explicit definitions of some of the terms we have been using informally, relying on examples and intuitions. Let us begin with a more precise characterization of a functional property:

> F is a *functional property* (or kind) just in case F can be characterized by a definition of the following form:
>
> For something *x* to have F (or to be an F) = $_{\text{def}}$ for *x* to have some property P such that C(P), where C(P) is a specification of the causal work that P is supposed to do in *x*.

We may call a definition having this form a "functional" definition. "C(P)," which specifies the causal role of F, is crucial. What makes a functional property the property it is, is the causal role associated with it; that is to say, F and G are the same functional property if and only if the causal role associated with F is the same as that associated with G. The term "causal work" in the

above schema of functional definitions should be understood broadly to refer to "passive" as well as "active" work: For example, if tissue damage causes P to instantiate in an organism, that is part of P's causal work or function. Thus, P's causal work refers to the *causal relations* involving the instances, or occurrences, of P in the organism or system in question.

Now we can define what it is for a property to "realize," or be a "realizer" of, a functional property:

> Let F be a functional property defined by a functional definition, as above. Property Q is said to *realize* F, or be a *realizer* or a *realization* of F, in system *x* if and only if C(Q), that is, Q fits the specification C in *x* (which is to say, Q in fact performs the specified causal work in system *x*).

Note that the definiens (the right-hand side) of a functional definition does not mention any particular property P that *x* has (when it has F); it only says that *x* has "some" property P fitting description C. In logical terminology, the definiens "existentially quantifies over" properties (it in effect says, "There exists some property P such that *x* has P and C[P]"). For this reason, functional properties are called "second-order" properties, with the properties quantified over (that is, properties eligible as instances of P) counting as "first-order" properties; they are second-order properties of a special kind—namely, those that are defined in terms of causal roles.

Let us see how this formal apparatus works. Consider the property of being a mousetrap. It is a functional property because it can be given the following functional definition:

> *x* is a mousetrap $=_{\text{def}}$ *x* has some property P such that P enables *x* to trap and hold or kill mice.

The definition does not specify any specific P that *x* must have; the causal work specified obviously can be done in many different ways. There are the familiar spring-loaded traps, and there are wire cages with a door that slams shut when a mouse enters; we can imagine high-tech traps with an optical sensor and all sorts of other devices. This means that there are many—in fact, indefinitely many—"realizers" of the property of being a mousetrap; that is, all sorts of physical mechanisms can be mousetraps.[14] The situation is the same with pain: A variety of physical/biological mechanisms can serve as tissue-damage detectors across biological species—and perhaps nonbiological systems as well.

Functionalism and Behaviorism

Both functionalism and behaviorism speak of sensory input and behavioral output—or "stimulus" and "response"—as central to the concept of mentality. In this respect, functionalism is part of a broadly behavioral approach to mentality and can be considered a generalized and more sophisticated version of behaviorism. But there are also significant differences between them, of which the following two are the most important.

First, the functionalist takes mental states to be *real internal* states of an organism with causal powers; for an organism to be in pain is for it to be in an internal state (for example, a neurobiological state for humans) that is typically caused by tissue damage and that in turn typically causes winces, groans, and avoidance behavior. And the presence of this internal state explains why humans react the way they do when they suffer tissue damage. In contrast, the behaviorist eschews talk of internal states entirely, identifying mental states with actual or possible behavior. Thus, to be in pain, for the behaviorist, is to wince and groan or be disposed to wince and groan, but not, as the functionalist would have it, to be in some *internal state that causes* winces and groans.

Although both the behaviorist and the functionalist may refer to "behavioral dispositions" in speaking of mental states, what they mean by "disposition" can be quite different: The functionalist takes a "realist" approach to dispositions, whereas the behaviorist embraces an "instrumentalist" line. We say that sugar cubes, for example, are soluble in water. But what does it mean to say that something is soluble in water? The answer depends on whether you adopt an instrumental or a realist view of dispositions. Let us see exactly how these two approaches differ:

> *Instrumentalist analysis*: x is soluble in water $=_{\text{def}}$ if x is immersed in water, x dissolves.

> *Realist analysis*: x is soluble in water $=_{\text{def}}$ x has an internal state S (for example, a certain microstructure) such that when x is immersed in water, S causes x to dissolve.

According to instrumentalism, therefore, all there is to the water solubility of a sugar is the fact that a certain conditional ("if-then") statement holds for it; thus, on this view, water solubility is a "conditional" or "hypothetical" property of the sugar cube—that is, the property of *dissolving if immersed in water*. Realism, in contrast, takes solubility to be a categorical, presumably microstruc-

tural, internal state of the cube of sugar that is causally responsible for its dissolving when placed in water. (Further investigation might reveal the state to be that of having a certain crystalline molecular structure.) Neither analysis requires the sugar cube to be placed in water or actually to be dissolving in order to be water-soluble. However, we may note the following difference: If *x* dissolves in water and *y* does not, the realist will give a causal explanation of this difference in terms of a difference in their microstructure. For the instrumentalist, the difference may just be a brute fact: It is just that the conditional "if placed in water, it dissolves" holds true for *x* but not for *y*, a difference that need not be grounded in any further differences between *x* and *y*.

In speaking of mental states as behavioral dispositions, then, the functionalist takes them as actual inner states of persons and other organisms that in normal circumstances cause behavior of some specific type under certain specified input conditions. Mental states serve as causal intermediaries between sensory input and behavioral output. In contrast, the behaviorist takes mental states merely as input-output, or stimulus-response, correlations. Many behaviorists (especially "radical" scientific behaviorists) believe that speaking of mental states as "inner causes" of behavior is scientifically unmotivated and philosophically unwarranted.[15]

The second significant difference between functionalism and behaviorism, one that gives the former a substantially greater theoretical power, is the way "input" and "output" are construed for mental states. For the behaviorist, input and output consist entirely of observable physical stimulus conditions and observable behavioral/physical responses. As mentioned earlier, the functionalist allows reference to other *mental states* in the characterization of a given mental state. It is a crucial part of the functionalist conception of mental states that their typical causes and effects can, and often do, include other mental states. Thus, for a ham sandwich to cause you to want to eat it, you must *believe* it to be a ham sandwich; a bad headache can cause you not only to frown and moan but also to experience further mental states like *distress* and a *desire* to call your doctor.

The two points that have just been reviewed are related: If you think of mental states as actual inner states of psychological subjects, you would regard them as having real causal powers, powers to cause and be caused by other states and events, and there is no obvious reason to exclude mental states from figuring among the causes or effects of other mental states. In conceiving mentality this way, the functionalist is espousing *mental realism*—a position that considers mental states as having a genuine ontological status and counts them among the phenomena of the world with a place in its causal structure. Mental

states are real for the behaviorist too, but only as behaviors or behavioral dispositions; for him, there is nothing mental over and above actual and possible behavior. For the functionalist, mental states are inner causes of behavior, and as such they are "over and above" behavior.

Including other mental events among the causes and effects of a given mental state is part of the functionalist's general conception of mental states as forming a complex causal network anchored to the external world at various points. At these points of contact, a psychological subject interacts with the outside world, receiving sensory inputs and emitting behavior outputs. And the identity of a given mental kind, whether it is a sensation like pain or a belief that it is going to rain or a desire for a ham sandwich, depends solely on the place it occupies in the causal network. That is, what makes a mental event the kind of mental event it is, is the way it is causally linked to other mental-event kinds and input-output conditions. Since each of these other mental-event kinds in turn has its identity determined by *its* causal relations to other mental events and to inputs and outputs, the identity of each mental kind depends ultimately on the whole system—its internal structure and the way it is causally linked to the external world via sensory inputs and behavior outputs. In this sense, functionalism gives us a *holistic* conception of mentality.

This holistic approach enables functionalism to sidestep one of the principal objections to behaviorism. This is the difficulty we saw earlier: A desire issues in overt behavior only when combined with an appropriate belief, and similarly, a belief leads to behavior only when a matching desire is present. For example, a person with a desire to eat an apple will eat an apple that is presented to her only if she believes it to be an apple (she would not bite into it if she thought it was a fake wooden apple); a person who believes that it is going to rain will take an umbrella only if she has a desire to stay dry. As we saw, this apparently makes it impossible to give a behavioral definition of desire without reference to belief or a definition of belief without reference to desire. The functionalist would say that this simply points to the holistic character of mental states: It is an essential feature of a desire that it is the kind of internal state that in concert with an appropriate belief causes a certain behavior output, and similarly for belief and other mental states.

But doesn't this make the definitions circular? If the concept of desire cannot be defined without reference to belief, and the concept of belief in turn cannot be explained without reference to desire, how can either be understood at all? We will see later (chapter 6) how the holistic approach of functionalism deals with this issue.[16]

Turing Machines

Functionalism was originally formulated by Putnam in terms of "Turing machines," mathematically characterized computing machines due to the British mathematician-logician Alan M. Turing.[17] Although it is now customary to formulate functionalism in terms of causal-functional roles—as we have done and will do in more detail in the next chapter—it is instructive to begin our systematic treatment of functionalism by examining the Turing-machine version of functionalism, usually called machine functionalism. This also gives us a background that will be helpful in exploring the idea that the workings of the mind are best understood in terms of the operations of a computing machine—that is, the computational view of the mind (computationalism, for short).

A Turing machine is made up of four components:

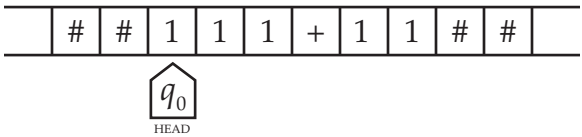
1. A *tape* divided into "squares" and unbounded in both directions
2. A *scanner-printer* ("head") positioned at one of the squares of the tape at any given time
3. A finite set of *internal states* (or *configurations*), $q_0, \ldots, q_n$
4. A finite *alphabet* consisting of symbols, $b_1, \ldots, b_m$

One and only one symbol appears on each square. (We may think of the blank as one of the symbols.)

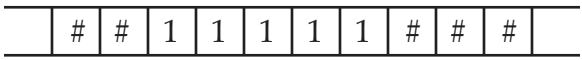
The machine operates in accordance with the following general rules:

A. At each time, the machine is in one of its internal states, q_i, and its head is scanning a particular square on the tape.
B. What the machine does at a given time t is completely determined by its internal state at t and the symbol its head is scanning at t.
C. Depending on its internal state and the symbol being scanned, the machine does three things:
   (1) Its head replaces the symbol with another (possibly the same) symbol of the alphabet. (To put it another way, the head erases the symbol being scanned and prints a new one, which may be the same as the erased one.)
   (2) Its head moves one square to the right or to the left (or halts, with the computation completed).
   (3) The machine enters into one of its internal states (which can be the same state).

Let us consider a Turing machine that adds positive integers in the unary notation. (In this notation, number n is represented as a sequence of n strokes, each stroke occupying one square.) Consider the following picture in which the problem of adding 3 and 2 is presented to the machine, which is to be started off with its head in state q_0 and scanning the first digit:



(The "scratch" symbol, #, marks the boundaries of the problem.) We want to "program" this Turing machine in such a way that when the computation is completed, the machine halts with a sequence of five consecutive strokes showing on the tape, like this:



It is easy to see that there are various procedures by which the machine could accomplish this. One simple way is to have the machine (or its head) move to the right looking for the symbol +, replace it with a stroke, keep moving right until it finds the right-most stroke, and when it does, erase it (that is, replace it with the scratch symbol #) and then halt. The following simple "machine table" is a complete set of instructions that defines our adder (call it TM_1):

	q_0	q_1
1	$1Rq_0$	#Halt
+	$1Rq_0$	
#	$\#Lq_1$	

Here is how we read this table. On the left-most column you find the symbols of the machine alphabet listed vertically, and the top row lists the machine's internal states. Each entry in the interior matrix is an *instruction*: It tells the machine what to do when it is scanning the symbol shown in the left-most column of that row and is in the internal state listed at the top of the column. For example, the entry $1Rq_0$, at the intersection of q_0 and 1, tells the machine: "If you are in internal state q_0 and scanning the symbol 1, replace 1 with 1 (that is, leave it unchanged), move to the right by one square, and go into internal state q_0 (that is, stay in the same state)." The entry immediately below, $1Rq_0$, tells the machine: "If you are in state q_0 and scanning the symbol +, replace + with 1, move to the

right by one square, and go into state q_0." The *L* in the bottom entry, $\#Lq_1$, means "move left by one square"; the entry in the right-most column, #Halt, means "If you are scanning 1 and in state q_1, replace 1 with # and halt." It is easy to see (the reader is asked to figure this out on her own) the exact sequence of steps our Turing machine will follow to compute the sum 3 + 2.

The machine table of a Turing machine is a complete and exhaustive specification of the machine's operations. We may therefore identify a Turing machine with its machine table. Since a machine table is nothing but a set of instructions, this means that a Turing machine can be identified with a set of such instructions.

What sort of things are the "internal states" of a Turing machine? We talk about this general question later, but with our machine TM_1, it can be helpful to think of the specific machine states in the following intuitive way: q_0 is a + and # searching state—it is a state such that when TM_1 is in it, it keeps going right, looking for + and #, ignoring any 1s it encounters. Moreover, if the machine is in q_0 and finds a +, it replaces it with a 1 and keeps moving to the right, while staying in the same state; when it scans a # (thereby recognizing the right-most boundary of the given problem), it backs up to the left and goes into a new state q_1, the "print # over 1 and then halt" state. When TM_1 is in this state, it will replace any 1 it scans with a # and halt. Thus, each state "disposes" the machine to do a set of specific things depending on the symbol being scanned (which therefore can be likened to sensory input).

But this is not the only Turing machine that can add numbers in unary notation; there is another one that is simpler and works faster. It is clear that to add unary numbers it is not necessary for the machine to determine the right-most boundary of the given problem; all it needs to do is to erase the initial 1 being scanned when it is started off, and then move to the right to find + and replace it with a 1. This is TM_2, with the following machine table:

	q_0	q_1
1	$\#Rq_1$	$1Rq_1$
+		1Halt
#		

We can readily build a third Turing machine, TM_3, that will do subtractions in the unary notation. Suppose the following subtraction problem is presented to the machine:

#	#	*b*	1	1	1	1	–	1	1	*b*	#	#

(Symbol *b* is used to mark the boundaries of the problem.) Starting the machine in state q_0 scanning the initial 1, we can write a machine table that computes *n*–*m* by operating like this:

1. The machine starts off scanning the first 1 of *n*. It goes to the right until it locates *m*, the number being subtracted. (How does it recognize it has located *m*?) It then erases the first 1 of this number (replacing it with a #), goes left, and erases the last 1 of *n* (again replacing it with a #).
2. The machine then goes right and repeats step 1 again and again, until it exhausts all the 1s in *m*. (How does the machine "know" that it has done this?) We then have the machine move right until it locates the subtraction sign–, which it erases (that is, replaces it with a #), and then halt. (If you like tidy output tapes, you may have the machine erase the *b*s before halting.)
3. If the machine runs out of the first set of strokes before it exhausts the second set (this means that $n < m$), we can have the machine print a certain symbol, say ?, to mean that the given problem is not well-defined. We must also provide for the case where $n = m$.

The reader is invited to write out a machine table that implements these operations.

We can also think of a "transcription machine," TM_4, that transcribes a given string of 1s to its right (or left). That is, if TM_4 is presented with the following tape to begin its computation,

	#	1	1	1	#	#	#	#	#	#	#	

it ends with the following configuration of symbols on its tape:

	#	1	1	1	#	1	1	1	#	#	#	#

The interest of the transcription machine lies in how it can be used to construct a multiplication machine, TM_5. The basic idea is simple: We can get $n \times m$ by transcribing the string of *n* 1s *m* times (that is, transcribing *n* repeatedly using *m* as a counter). The reader is encouraged to write a machine table for TM_5.

Since any arithmetical operation (squaring, taking the factorial, and so on) on natural numbers can be defined in terms of addition and multiplication,

it follows that there is a Turing machine that computes any arithmetical operation. More generally, it can be shown that any computation performed by any computer can be done by a Turing machine. That is, being computable and being computable by a Turing machine turn out to be equivalent.[18] In this sense, the Turing machine captures the general idea of computation and computability.

We can think of a Turing machine with two separate tapes (one for input, on which the problem to be computed is presented, and the other for actual computation and the final output) and two separate heads (one for scanning and one for printing). This helps us to think of a Turing machine as receiving "sensory stimuli" (the symbols on the input tape) through its scanner ("sense organ") and emitting specific behaviors in response (the symbols printed on the output tape by its printer head). It can be shown that any computation that can be done by a two-tape machine or a machine with any finite number of tapes can be done by a one-tape machine. So adding more tapes does not strengthen the computing power of Turing machines or substantively enrich the concept of a Turing machine, although it could speed up computations.

Turing also showed how to build a "universal machine," which is like a general-purpose computer in that it is not dedicated to the computation of a specific function but can be programmed to compute any function you want. On the input tape of this machine, you specify two things: the machine table of the desired function in some standard notation that can be read by the universal machine and the values for which the function is to be computed. The universal machine is programmed to read any machine table and carry out the computation in accordance with the instructions of the machine table.

The notion of a Turing machine can be generalized to yield the notion of a *probabilistic automaton*. As you recall, each instruction of a Turing machine is *deterministic*: Given the internal state and the symbol being scanned, the immediate next operation is wholly and uniquely determined. An instruction of a probabilistic, or stochastic, automaton has the following general form: Given internal state *qi* and scanned symbol *bj*:

1. Print b_k with probability r_1, or print b_l with probability r_2, . . . , or print b_m with probability r_n (where the probabilities add up to 1).
2. Move *R* with probability r_1, or move *L* with probability r_2 (where the probabilities add up to 1).
3. Go into internal state q_j with probability r_1, or into q_k with probability r_2, . . . , or into q_m with probability r_n (again, the probabilities adding up to 1).

Although in theory a machine can be made probabilistic along any one or more of these three dimensions, it is customary to understand a probabilistic machine as one that incorporates probabilities into state transitions, in the manner of (3) above. The operations of a probabilistic automaton are not deterministic; the current internal state of the machine and the symbol it is scanning do not—do not always, at any rate—together uniquely determine what the machine will do next. However, the behavior of such a machine is not random or arbitrary either: There are fixed and stable probabilities describing the machine's operations. If we are thinking of a machine that describes the behavior of an actual psychological subject, a probabilistic machine may be more realistic than a deterministic one; however, we may note that it is generally possible to construct a deterministic machine that simulates the behavior of a probabilistic machine to any desired degree of accuracy, which makes probabilistic machines theoretically dispensable.

Physical Realizers of Turing Machines

Suppose that we give the machine table for our simple adding machine, TM_1, to an engineering class as an assignment: Each student is to build an actual physical device that will do the computations as specified by its machine table. What we are asking the students to build, therefore, are "physical realizers" of TM_1—real-life physical computing machines that will operate in accordance with the machine table of TM_1. We can safely predict that a huge, heterogeneous variety of machines will be turned in. Some of them may really look and work like the Turing machine as described: They will have a paper tape neatly divided into squares, with an actual physical "head" that can read, erase, and print symbols. Some will perhaps use magnetic tapes and heads that read, write, and erase electrically. Some machines will have no "tapes" or "heads" but instead use spaces on a computer disk or memory locations in their CPU to do the computation. A clever student with a sense of humor (and lots of time and other resources) might try to build a hydraulically operated device with pipes and valves instead of wires and switches. The possibilities are endless.

But what exactly is a physical realizer of a Turing machine? What makes a physical device a *realizer of* a given Turing machine? First, the symbols of the machine's alphabet must be given concrete physical embodiments; they could be blotches of ink on paper, patterns of magnetized iron particles on plastic tape, electric charges in capacitors, or what have you. Whatever they are, the physical device that does the "scanning" must be able to "read" them—that is,

differentially respond to them—with a high degree of reliability. This means that the physical properties of the symbols place a set of constraints on the physical design of the scanner, but these constraints need not, and usually will not, determine a unique design; a great multitude of physical devices are likely to be adequate to serve as a scanner for any set of physically embodied symbols. The same considerations apply to the machine's printer and outputs as well: The symbols the machine prints on its output tape (we are thinking of a two-tape machine) must be given physical shapes, and the printer must be designed to produce them on demand. The printer, of course, does not have to "print" anything in a literal sense; the operation could be wholly electronic, or the printer could be a speaker that vocalizes the output or an LCD monitor that visually displays it (and saves it for future computational purposes).

What about the "internal states" of the machine? How are they physically realized? Consider a particular instruction on the machine table of TM_1: If the machine is in state q_0 and scanning a +, replace the + with a 1, move right, and go into state q_1. Assume that Q_0 and Q_1 are the physical states realizing q_0 and q_1, respectively. Q_0 and Q_1, then, must satisfy the following condition: An occurrence of Q_0, together with the physical scanning of +, must *physically cause* three physical events: (1) The physical symbol + is replaced with the physical symbol 1; (2) the physical scanner-printer (head) moves one square to the right (on the physical tape) and scans it; and (3) the machine enters state Q_1. In general, then, what needs to be done is to *replace the functional or computational relations* among the various abstract parameters (symbols, states, and motions of the head) mentioned in the machine table *with matching causal relations among the physical embodiments* of these parameters. That is to say, a physical realizer of a Turing machine is a physical causal mechanism that is isomorphic to the machine table of the Turing machine.

From the logical point of view, the internal states are only "implicitly defined" in terms of their relations to other parameters: q_j is a state such that if the machine is in it and scanning symbol b_k, the machine replaces b_k with b_l, moves R (that is, to the right), and goes into state q_h; if the machine is scanning b_m, it does such and such; and so on. So q_j can be thought of as a function that maps symbols of the alphabet to the triples of the form $<b_k$, R (or L), $q_h>$. From the physical standpoint, Q_j, which realizes q_j, can be thought of as a *causal* intermediary between the physically realized symbols and the physical realizers of the triples—or equivalently, as a *disposition* to emit appropriate physical outputs (the triples) in response to different physical stimuli (the physical symbols scanned). This means that the intrinsic physical natures of the Qs that realize the *q*s are of no interest to us as long as they have the right

causal powers or capacities; their intrinsic properties do not matter—or more accurately, they matter only to the extent that they affect the desired causal powers of the states and objects that have them. As long as these states perform their assigned causal work, they can be anything you please. Clearly, whether the Qs realize the *q*s depends crucially on how the tape, symbols, and so on are physically realized; in fact, these are interdependent questions. It is plausible to suppose that, with some mechanical ingenuity, a machine could be rewired so that physical states realizing distinct machine states could be interchanged without affecting the operation of the machine.

We see, then, a convergence of two ideas: the functionalist conception of a mental state as a state occupying a certain specific causal role and the idea of a physical state realizing an internal state of a Turing machine. Just as, on the functionalist view, what makes a given mental state the kind of mental state it is, is its causal role with respect to sensory inputs, behavior outputs, and other mental states, so what makes a physical state the realizer of a given internal machine state is its causal relations to inputs, outputs, and other physical realizers of the machine's internal states. This is why it is natural for functionalists to look to Turing machines for a model of the mind.

Let S be a physical system (which may be an electromechanical device like a computer, a biological organism, an auto assembly plant, or anything else), and assume that we have adopted a vocabulary to describe its inputs and outputs. That is, we have a specification of what is to count as the inputs it receives from its surroundings and what is to count as its behavioral outputs. Assume, moreover, that we have specified what states of S are to count as its "internal states." We will say that a Turing machine M is a *machine description* of system S, relative to a given input-output specification and a specification of the internal states, just in case S realizes M relative to the input-output and internal state specifications. Thus, the relation of *being a machine description of* is the converse of the relation of *being a realizer (or realization) of*. We can also define a concept that is weaker than machine description: Let us say that a Turing machine M is a *behavioral description* of S (relative to an input-output specification) just in case M provides a correct description of S's input-output correlations. Thus, every machine description of S is also a behavioral description of S, but the converse does not in general hold. M can give a true description of the input-output relations characterizing S, but its machine states may not be realized in S, and S's inner workings (that is, its computational processes) may not correctly mirror the functional-computational relationships given by M's machine table. In fact, there may be another Turing machine M*, distinct from M, that gives a correct machine description of S. It follows, then,

that *two physical systems that are input-output equivalent may not be realizations of the same Turing machine*. (The pair of adding machines TM_1 and TM_2 is a simple example of this.)

Machine Functionalism: Motivations and Claims

Machine functionalists claim that we can think of the mind as a Turing machine (or a probabilistic automaton). This of course needs to be filled out, but from the preceding discussion it should be pretty clear how the story will go. The central idea is that what it is for something to have mentality—that is, to have a psychology—is for it to be a physically realized Turing machine of appropriate complexity, with its mental states (that is, mental-state types) identified with the realizers of the internal states of the machine table. Another way of explaining this idea is to use the notion of machine description: An organism has mentality just in case there is a Turing machine of appropriate complexity that is a machine description of it, and its mental-state kinds are to be identified with the physically realized internal states of that Turing machine. All this is, of course, relative to an appropriately chosen input-output specification, since you must know, or decide, what is to count as the organism's inputs and outputs before you can determine what Turing machine (or machines) it can be said to realize.

Let us consider the idea that *the psychology of an organism* can be represented by a Turing machine, an idea that is commonly held by machine functionalists.[19] Let V be a complete specification of all possible inputs and outputs of a psychological subject S, and let C be all actual and possible input-output correlations of S (that is, C is a complete specification of which input applied to S elicits which output, for all inputs and outputs listed in V). In constructing a *psychology* for S, we are trying to formulate a *theory* that gives a perspicuous systematization of C by positing a set of internal states in S. Such a theory *predicts* for any input applied to S what output will be emitted by S and also *explains* why that particular input will elicit that particular output. It is reasonable to suppose that for any behavioral system complex enough to have a psychology, this kind of systematization is not possible unless we advert to its internal states, for we must expect that the same input applied to S does not always prompt S to produce the same output. The actual output elicited by a given input depends, we must suppose, on the internal state of S at that time.

Before we proceed further, it is necessary to modify our notion of a Turing machine in one respect: The internal states, *q*s, of a Turing machine are *total*

states of the machine at a given time, and the Qs that are their physical realizers are also *total* physical states at a time of the physically realized machine. This means that the Turing machines we are talking about are not going to look very much like the psychological theories we are familiar with; the states posited by these theories are seldom, if ever, total states of a subject at a time. But this is a technical problem, something we assume can be remedied with a more fine-grained notion of an "internal state." We can then think of a total internal state as made up of these "partial" states, which combine in different ways to yield different total states. This modification should not change anything essential in the original conception of a Turing machine. In the discussion to follow, we use this modified notion of an internal state in most contexts.

To return to the question of representing the psychology of a subject S in terms of a Turing machine: What Turing machine, or machines, is adequate as a description of S's psychology? Evidently, any adequate Turing machine must be a behavioral description of S, in the sense defined earlier; that is, it must give a correct description of S's input-output relations (relative to V). But as we have seen, there is bound to be more than one Turing machine—in fact, if there is one, there will be indefinitely more—that gives a correct representation of S's input-output relations.

Since each of these machines is a correct behavioral description of our psychological subject S, they are all equally good as *predictive* theories: Although some of them may be easier to manipulate and computationally more efficient than others, they all predict the same behavior output for the same input. This is a simple consequence of the notion of "behavioral description." But they are different as Turing machines. But do the differences between them matter?

It should be clear how behaviorally equivalent Turing machines, say, M_1 and M_2, can differ from each other. To say that they are different Turing machines is to say that their machine tables are different—that is how Turing machines are individuated. This means that when they are given the same input, M_1 and M_2 are likely to go through *different computational processes* to arrive at the same output. Each machine has a set of internal states—let us say $<q_0, q_1, \ldots, q_n>$ for M_1 and $<r_0, r_1, \ldots, r_m>$ for M_2. Let us suppose further that M_1 is a machine description of our psychological subject S, but M_2 is not. That is, S is a physical realizer of M_1 but not of M_2. This means that the computational relations represented in M_1, but not those represented in M_2, are mirrored in a set of causal relations among the physical-psychological states of S. So there are real physical (perhaps neurobiological) states in S, $<Q_0, Q_1, \ldots, Q_n>$, corresponding to M_1's internal states $<q_0, q_1, \ldots, q_n>$, and these

Qs are causally hooked up to each other and to the physical scanner (sense organs) and the physical printer (motor mechanisms) in a way that ensures that for all computational processes generated by M_1, isomorphic causal processes occur in S. As we may say, S is a "causal isomorph" of M_1.

There is, then, a clear sense in which M_1 is, but M_2 is not, *psychologically real* for S, even though they are both accurate predictive theories of S's observable input-output behaviors. M_1 gives "the true psychology" of S in that, as we saw, S has a physical structure whose states constitute a causal system that mirrors the computational structure represented by the machine table of M_1, and the physical-causal operations of S form an isomorphic image of the computational operations of M_1. This makes a crucial difference when what we want is an *explanatory* theory, a theory that *explains why, and how, S does what it does under the given input conditions*. Suppose we say: When input *i* was applied to S, S emitted behavioral output *o because* it was in internal state Q. This can count as an explanation, it seems, only if the state appealed to—namely, Q—is a "real" state of the system. In particular, it can count as a *causal* explanation only if the state Q is what, in conjunction with *i*, caused *o*. Since S is a physical realizer of M_1, or equivalently, M_1 is a machine description of S, the causal process leading from Q and input *i* to behavior output *o* is mirrored exactly by the computational process that occurs in accordance with the machine table of M_1. In contrast, Turing machine M_2, which is not realized by S, has no "inner" psychological reality for S, even though it correctly captures all of S's input-output connections. Although, like M_1, M_2 correlates input *i* with output *o*, the computational process whereby the correlation is effected does not reflect the actual causal process in S that leads from *i* to *o* (or physical embodiments thereof). The explanatory force of "because" in "S emitted *o* when it received input *i because* it was in state Q" derives from the causal relations involving Q and the physical embodiments of *o* and *i* in the system S.

The philosophical issues here depend, partly but critically, on the metaphysics of scientific theories you accept. If you think of scientific theories in general, or theories over some specific domain, merely as predictive instruments that enable us to infer or calculate further observations from the given data, you need not attach any existential significance to the posits of these theories—like the unobservable microparticles of theoretical physics and their (often quite strange) properties—and may regard them only as calculational aids in deriving predictions. A position like this is called "instrumentalism," or "antirealism," about scientific theory.[20] On such a view, the issue of "truth" does not arise for the theoretical principles, nor does the issue of "reality" for the entities and properties posited; the only thing that matters is the

"empirical, or predictive, adequacy" of the theory—how accurately the theory works as a predictive device and how comprehensive its coverage is. If you accept an instrumentalist stance toward psychological theory, therefore, any Turing machine that is a behavioral description of a psychological subject is good enough, exactly as good as any other behaviorally adequate description of it; you may prefer some over others on account of manipulative ease and computational cost, but the question of "reality" or "truth" does not arise. If this is your view of the nature of psychology, you will dismiss as meaningless the question which of the many behaviorally adequate psychologies is "really true" of the subject.

But if you adopt the perspective of "realism" on scientific theories, or at any rate about psychology, you will not think all behaviorally adequate descriptions are psychologically adequate. An adequate psychology for the realist must have "psychological reality": That is, the internal states it posits must be the real states of the organism with an active role as causal intermediaries between sensory inputs and behavior outputs, and this means that only a Turing machine that is a correct machine description of the organism is an acceptable psychological theory. The simplest and most elegant behavioral description may not be the one that correctly describes the inner processes that cause the subject's observable behavior; there is no a priori reason to suppose that our subject is put together according to the specifications of the simplest and most elegant theory (whatever your standards of simplicity and elegance might be).

Why should one want to go beyond the instrumentalist position and insist on psychological reality? There are two related reasons: (1) Psychological states, namely, the internal states of the psychological subject posited by a psychology, must be regarded as real, as we saw, if we expect the theory to generate explanations, especially causal explanations, of behavior. And this seems to be the attitude of working psychologists: It is their common, almost universal, practice to attribute to their subjects internal states, capacities, functions, and mechanisms (for example, information processing and storage, reasoning and inference, mental imagery, preference structures) and to refer to them in formulating what they regard as causal explanations of behavior. Further, (2) it seems natural to expect—this seems true of most psychologists and cognitive scientists—to find actual neural-biological mechanisms that underlie the psychological states, capacities, and functions posited by correct psychological theories. Research in the neural sciences, in particular cognitive neuroscience, have had impressive successes—and we expect this to continue—in identifying physiological mechanisms that implement psychological and cognitive capacities and functions. It is a reflection of our realistic stance toward psychological

theorizing that we generally expect, and sometimes insist on, physiological foundations for psychological theories. The requirement that the correct psychology of an organism be a machine description of it,[21] not merely a behaviorally adequate one, can be seen as an expression of a commitment to realism about psychological theory.

If the psychology of any organism can be represented as a Turing machine, it is natural to consider the possibility of using representability by a Turing machine to explicate, or define, what it is for something to have a psychology. As we saw, that precisely is what machine functionalism proposes: What it is for an organism, or system, to have a psychology—that is, what it is for an organism to have mentality—is for it to realize an appropriate Turing machine. It is not merely that anything with mentality has an appropriate machine description; machine functionalism makes the stronger claim that its having a machine description of an appropriate kind is *constitutive of* its mentality. This is a philosophical thesis about the nature of mentality: Mentality, or having a mind, consists in being a physical computer that realizes a Turing machine of appropriate complexity and powers. What makes us creatures with mentality, therefore, is the fact that we are Turing machines. Having a brain is important to mentality, but the importance of the brain lies exactly in its being a computing machine. It is our brain's computational powers, not its biological properties and functions, that constitute our mentality. In short, our brain is our mind because it is a computing machine, not because it is composed of the kind of protein-based biological stuff it is composed of.

Machine Functionalism: Further Issues

Suppose that two systems, S_1 and S_2, are *in the same mental state* (at the same time or different times). What does this mean on the machine-functionalist conception of a mental kind? A mental kind, as you will remember, is supposed to be an internal state of a Turing machine (of an "appropriate kind"); so for S_1 and S_2 to be in the same state, there must be some Turing machine state q such that S_1 is in q and S_2 is also in q. But what does this mean?

S_1 and S_2 are both physical systems, and we know that they could be systems of very different sorts (recall multiple realizability). As physical systems, they have physical states (that is, they instantiate certain physical properties); to say that they are both in machine state q at time t is to say this: There are physical states Q_1 and Q_2 such that Q_1 realizes q in S_1, and Q_2 realizes q in S_2, and, at t, S_1 is in Q_1 and S_2 in Q_2. Multiple realizability tells us that Q_1 and Q_2 need not have much in common qua physical states; one could be a biological

state and the other an electronic one. What binds the two states together is only the fact that in their respective systems they implement the same internal machine state. That is to say, the two states play the same computational role in their respective systems.

But talk of "the same internal machine state q" makes sense only in relation to a specific machine table. That is to say, internal states of a Turing machine are identifiable only relative to a particular machine table: In terms of the layout of machine tables we used earlier, an internal state q is wholly characterized by the vertical column of instructions appearing under it. But these instructions refer to other internal states, say, q_i, q_j, and q_k, and if you look up the instructions falling under these, you are likely to find references back to state q. So these states are inter-defined. What all this means is that *the sameness or difference of an internal state across different machine tables—that is, across different Turing machines—has no meaning*. It makes no sense to say of an internal state q_i of one Turing machine and a state q_k of another Turing machine that q_i is, or is not, the same state as q_k; nor does it make sense to say of a physical state Q_i of a physically realized Turing machine that it realizes, or does not realize, the same internal machine state q as does a physical state Q_k of another physical machine, *unless the two physical machines are realizations of the same Turing machine*.

Evidently, then, the machine-functionalist conception of mental kinds has the following consequence: For any two subjects to be in the same mental state, they must realize the same Turing machine. But if they realize the same Turing machine, their total psychology must be identical. That is, on machine functionalism, two subjects' total psychology must be identical if they are to share even a single psychological state—or even to give meaning to the talk of their being, or not being, in the same psychological state. This sounds absurd: It does not seem reasonable to require that for two persons to share a mental state—say, the belief that snow is white—the total set of psychological regularities governing their behavior must be exactly identical. Before we discuss this issue further, we must attend to another matter, and this is the problem of how the inputs and outputs of a system are to be specified.

Suppose that two systems, S_1 and S_2, realize the same Turing machine; that is, the same Turing machine gives a correct machine description for each. We know that realization is relative to a particular input-output specification; that is, we must know what is to count as input conditions and what is to count as behavior outputs before we can tell whether it realizes a given Turing machine. Let V_1 and V_2 be the input-output specifications for S_1 and S_2, respectively, relative to which they realize the same Turing machine. Since the same machine

table is involved, V_1 and V_2 must be isomorphic: The elements of V_1 can be correlated, one to one, with the elements of V_2 in a way that preserves their roles in the machine table.

But suppose that S_1 is a real psychological system, perhaps a human (call him Larry), whereas S_2 is a computer, an electromechanical device (call it MAX). So the inputs and outputs specified by V_2 are the usual inputs and outputs appropriate for a computing machine, perhaps strings of symbols entered on the keyboard or images scanned by a video camera as input and symbols or images displayed on the monitor or its printout as output. According to machine functionalism, Larry and MAX have the same psychology. But shouldn't this strike us as absurd? One might say: MAX is only a computer simulation of Larry's psychology, and in granting MAX the full psychological status that we grant Larry, machine functionalism is *conflating a psychological subject with a computer simulation of it*. No one will confuse the operation of a jet engine or the spread of rabies in wildlife with their computer simulations. It is difficult to believe that this distinction suddenly vanishes when we perform a computer simulation of the psychology of a person. (We will return to this question below in a section on computationalism and the Chinese room argument.)

One thing that obviously seems wrong about our computer, MAX, as a psychological system when we compare it with Larry is its inputs and outputs: Although its input-output specification is isomorphic to Larry's, it seems entirely inappropriate for psychology. It may not be easy to characterize the differences precisely, but we would not consider inputs and outputs consisting merely of strings of symbols, or electronic images, as appropriate for something with true mentality. Grinding out strings of symbols is not like the full-blown behavior that we see in Larry. For one thing, MAX's outputs have nothing to do with its survival or continued proper functioning, and its inputs do not have the function of providing MAX with information about its surroundings. As a result, MAX's outputs lack what may be called "teleological aptness" as a response to its inputs. All this makes it difficult to think of MAX's outputs as constituting real behavior or action, something that is necessary if we are to regard it as a genuine psychological system.

Qua realizations of a Turing machine, MAX and Larry are symmetrically related. If, however, we see here an asymmetry in point of mentality, it is clear that the nature of inputs and outputs is an important factor, and our considerations seem to show that for a system realizing a Turing machine to count as a psychological system, its input-output specification (relative to which it realizes the machine) must be *psychologically appropriate*. Exactly what this appropriateness consists in is an interesting and complex question that requires

further exploration. In any case, the machine functionalist must confront this question: Is it possible to give a characterization of this input-output appropriateness that is consistent with functionalism—in particular, without using mentalistic terms or concepts? Recall a similar point we discussed in connection with behaviorism: Not to beg the question, the behavior that the behaviorist is allowed to talk about in giving behavioristic definitions of mental concepts must be "physical behavior," not intentional action with an explicit or implicit mental component (such as reading the morning paper, being rude to a waiter, or going to a music concert). If your project is to get mentality out of behavior, your notion of behavior must not presuppose mentality.

The same consideration applies to the machine functionalist: Her project is to define mentality in terms of Turing machines and input-output relations. The additional tool she can make use of, something not available to the behaviorist, is the concept of a Turing machine with its "internal" states, but her input and output are subject to the same constraint—her input-output, like the behaviorist's, must be physical input-output. If this is right, it seems no easy task for the machine functionalist to distinguish, in a principled way, Larry's inputs-outputs from MAX's, and hence genuine psychological systems from their simulations. We pointed out earlier that Larry's outputs, given his inputs, seem *teleologically apt*, whereas MAX's do not. They have something to do with his proper functioning in his environment—coping with the ever-changing conditions of his surroundings and satisfying his needs and desires. But can this notion of teleology—purposiveness or goal-directedness—be explained in a psychologically neutral way, without begging the question? Perhaps some biological-evolutionary story could be attempted, but it remains an open question whether such a bioteleological program will succeed. These considerations give credence to the idea that in order to have genuine mentality, a system must be embedded in a natural environment (ideally including other systems like it), interacting and coping with it and behaving appropriately in response to the new, and changing, conditions it encounters.

Let us now return to the question of whether machine functionalism is committed to the consequence that two psychological subjects can share a psychological state only if they have an identical total psychology. As we saw, the implication follows from the fact that, on machine functionalism, being in the same psychological state is being in the same internal machine state and that the sameness, or difference, of machine states makes sense only in relation to the same Turing machine, and never across distinct Turing machines. What is perhaps worse, it also follows that it makes no sense to say that two psychological subjects are *not* in the same psychological state unless they have

an identical total psychology! But this conclusion must be slightly weakened in consideration of the fact that the input-output specifications of the two subjects realizing the same Turing machine may be different and that the individuation of psychologies may have to be made sensitive to input-output specifications (we return shortly to this point). So let us speak of "isomorphic" psychologies for psychologies that are instances of the same Turing machine *modulo* input-output specification. We then have the following result: On machine functionalism, for two psychological subjects to share even a single mental state, their total psychologies must be isomorphic to each other. Recall Putnam's complaint against the psychoneural identity theory: This theory makes it impossible for both humans and octopuses to be in the same pain state unless they share the same brain state, an unlikely possibility. But we now see that the table is turned against Putnam's machine functionalism: For an octopus and a human to be in the same pain state, they must share an isomorphic psychology—an unlikely possibility, to say the least! And for two humans to share a single mental state, they must have an identical total psychology (since the same input-output specification presumably must hold for all or most humans). No analogous consequence follows from the psychoneural identity theory; in this respect, therefore, machine functionalism seems to fare worse than the theory it hopes to replace. All this is a consequence of a fact mentioned earlier, namely, that on functionalism, the individuation of mental kinds is essentially *holistic*; that is, what makes a given mental kind the kind it is depends on its relationships to other mental kinds, where the identities of these other mental kinds depend similarly on their relationships to still other mental kinds, and so on.

Things are perhaps not as bleak for machine functionalism, however, as they might appear, for the following line of response seems available: For both humans and octopuses to be in pain, it is not necessary that *total* octopus psychology coincide with, or be isomorphic to, *total* human psychology. It is only necessary that there be *some* Turing machine that is a correct machine description of both and in which pain figures as an internal machine state; it does not matter if this shared Turing machine falls short of the maximally detailed Turing machines that describe them (these machines represent their "total psychologies"). So what is necessary is that humans and octopuses share a partial, or abbreviated, psychology that covers pains (and perhaps also related sensations). Whether "pain psychology" can be so readily isolated, or abstracted, from a total psychology is a question worth pondering, especially in the context of the functionalist conception of mentality, but there is another related issue that we should briefly consider.

Recall the point that all this talk of humans' and octopuses' realizing a Turing machine is relative to an input-output specification. Doesn't this mean, in view of our earlier discussion of a real psychological subject and a computer simulation of one, that the input and output conditions characteristic of humans when they are in pain must be appropriately similar, if not identical, to those characteristic of octopuses' pains, if both humans and octopuses can be said to be in pain? Consider the output side: Do octopuses wince and groan in reaction to pain? They perhaps can wince, but they surely cannot groan or scream and yell "Ouch!" How similar is octopuses' escape behavior, from the purely physical point of view, to the escape behavior of, say, middle-aged, middle-class American males? Is there an abstract enough *nonmental* description of pain behavior that is appropriate for humans and octopuses and all other pain-capable organisms and systems? If there is not, machine functionalism seems to succumb again to the same difficulty that the functionalist has charged against the brain-state theory: An octopus and a human cannot be in the same pain state. Again, the best bet for the functionalist seems to be to appeal to the "teleological appropriateness" of an octopus's and a person's escape behaviors—that is, the fact that the behaviors are biologically appropriate responses to the stimulus conditions in enhancing their chances of survival and their well-being in their respective environments.

There is a further "appropriateness" issue for Turing machines that we must raise at this point. You will remember our saying that for a machine functionalist, a system has mentality just in case it realizes an "appropriately complex" Turing machine. This proviso is necessary because there are all sorts of simple Turing machines (recall our sample machines) that clearly do not suffice to generate mentality. But how complex is complex enough? What is complexity anyway, and why does it matter? And what kind of complexity is "appropriate" for mentality? These are important but difficult questions, and machine functionalism, unsurprisingly, has not produced detailed general answers to them. What we have, though, is an intriguing proposal, from Alan Turing himself, of a test to determine whether a computing machine can "think." This is the celebrated "Turing test," and this is the right time to consider Turing's proposal.

Can Machines Think? The Turing Test

Turing's innovative proposal is to bypass these general theoretical questions about appropriateness in favor of a concrete operational test that can evaluate the performance capabilities of computing machines vis-à-vis average hu-

mans who, as all sides would agree, are fully mental.[22] The idea is that if machines can do as well as humans on certain cognitive, intellectual tasks, then they must be judged no less psychological ("intelligent") than humans. What, then, are these tasks? Obviously, they must be those that, intuitively, require intelligence and mentality to perform. Turing describes a game, the "imitation game," to test for the presence of these capacities.

The imitation game is played as follows. There are three players: the interrogator, a man, and a woman, with the interrogator segregated from the other two in another room. The man and woman are known only as "X" and "Y" to the interrogator, whose object is to identify which is the man and which is the woman by asking questions via keyboard terminals and monitors. The man's object is to mislead the interrogator to make an erroneous identification, whereas the woman's job is to help the interrogator. There are no restrictions on the topics of the questions asked.

Suppose, Turing says, we now replace the man with a computing machine. The machine is programmed to simulate the part played by the man to fool the interrogator into making wrong guesses. Will the machine do as well as the man in fooling the interrogator? Turing's proposal is that if the machine does as well as the man, then we must credit it with all the intelligence that we would normally confer on a human; it must be judged to possess the full mentality that humans possess.[23]

The gist of Turing's idea can be captured in a simpler test: By asking questions (or just holding a conversation) via keyboard terminals, can we find out whether we are conversing with a human or a computing machine? (This is the way the Turing test is now being conducted.) If a computer can consistently fool us so that our success in ascertaining its identity is no better than what could be achieved by random guesses, we must concede, it seems, that this machine has the kind of mentality that we grant to humans. There already are chess-playing computers that would fool most people this way, but only in playing chess: Average chess players would not be able to tell if they are playing a human opponent or a computer. But the Turing test covers all possible areas of human concern: politics and sports, music and poetry, how to fix a leaking faucet or make a soufflé—no holds are barred.

The Turing test is designed to isolate the questions of intelligence and mentality from irrelevant considerations, such as the appearance of the machine (as Turing points out, it does not have to win beauty contests), details of its composition and structure, whether it speaks and moves about like a human, and so on. The test is to focus on a broad range of rational, intellectual capacities and functions. But how good is the test?

Some have objected that the test is too tough and too narrow. Too tough because something does not have to be smart enough to outwit a human to have mentality or intelligence; in particular, the possession of a language should not be a prerequisite for mentality (think of mute animals). Human intelligence itself encompasses a pretty broad range, and there appears to be no compelling reason to set the minimal threshold of mentality at the level of performance required by the Turing test. The test is perhaps also too narrow in that it seems at best to be a test for the presence of *humanlike* mentality, the kind of intelligence that characterizes humans. Why couldn't there be creatures, or machines, that are intelligent and have a psychology but would fail the Turing test, which, after all, is designed to test whether the computer can fool a *human* interrogator into thinking it is a *human*? Furthermore, it is difficult to see it as a test for the presence of mental states like sensations and perceptions, although it may be a good test of broadly intellectual and cognitive capacities (reasoning, memory, and so on). To see something as a full psychological system we must see it in a real-life context, we might argue; we must see it coping with its environment, receiving sensory information from its surroundings, and behaving appropriately in response to it.

Various replies can be attempted to counter these criticisms, but can we say that the Turing test at least provides us with a *sufficient* condition for mentality, although, for the reasons just stated, it cannot be considered a *necessary* condition? If something passes the test, it is at least as smart as we are, and since we have intelligence and mentality, it would be only fair to grant it the same status—or so we might argue. This reasoning seems to presuppose the following thesis:

> *Turing's Thesis*. If two systems are input-output equivalent, they have the same psychological status; in particular, one is mental, or intelligent, just in case the other is.

We call it Turing's Thesis because Turing appears to be committed to it. Why? Because the Turing test looks only at inputs and outputs: If two computers produce the same output for the same input, for all possible inputs—that is, if they are input-output equivalent—their performance on the Turing test will be exactly identical,[24] and one will be judged to have mentality if and only if the other is. This means that if two Turing machines are correct behavioral descriptions of some system (relative to the same input-output specification), they are psychological systems to the same degree. In this way the general philosophical stance implicit in Turing's Thesis is more behavioristic than

machine-functionalist. For machine functionalism is consistent with the denial of Turing's thesis: It says that input-output equivalence, or behavioral equivalence, is not sufficient to guarantee the same degree of mentality. What follows from machine functionalism is only that systems that realize the same Turing machine—that is, systems for which an identical Turing machine is a correct machine description—enjoy the same degree of mentality.

It appears, then, that Turing's Thesis is mistaken: Internal processing ought to make a difference to mentality. Imagine two machines, each of which does basic arithmetic operations for integers up to 100. Both give correct answers for any input of the form $n + m$, $n \times m$, $n–m$, and $n \div m$ for whole numbers n and m less than or equal to 100. But one of the machines calculates ("figures out") the answer by applying the usual algorithms we use for these operations, whereas the other has a file in which answers are stored for all possible problems of addition, multiplication, subtraction, and division for integers up to 100, and its computation consists in "looking up" the answer for any problem given to it. The second machine is really more like a filing system than a computing machine; it does nothing that we would normally describe as "calculation" or "computation." Neither machine is nearly complex enough to be considered for possible mentality; however, the example should convince us that we need to consider the structure of internal processing, as well as input-output correlations, in deciding whether a given system has mentality.[25] If this is correct, it shows the inadequacy of a purely behavioral test, such as the Turing test, as a criterion of mentality.

So Turing's Thesis seems incorrect: Input-output equivalence does not imply equal mentality. But this does not necessarily invalidate the Turing test, for it may well be that given the inherent richness and complexity of the imitation game, any computing machine that can consistently fool humans—in fact, any machine that is in the ballpark for the competition—has to be running a highly sophisticated, unquestionably "intelligent" program, and there is no real chance that this machine could be operating like a gigantic filing system with a superfast retrieval mechanism.[26] We should note that the computing machines' performance at actual Turing tests—and these have been restricted tests, on specific topics—has been truly dismal so far; computers programmed to fool human judges have not come anywhere near their goal. Turing's prediction in 1950 that in fifty years we would see computers passing his test has missed the mark—by a huge margin. It is also true, though, that designing a "thinking" machine that will pass the Turing test has not been a priority for artificial-intelligence researchers for the past several decades.

Computationalism and the "Chinese Room"

Computationalism, or the computational theory of mind, is the view that cognition, human or otherwise, is information processing, and that information processing is computation over symbolic representations according to syntactic rules, rules that are sensitive only to the shapes of these representations. This view of mental, or cognitive, processes, which arguably is the reigning research paradigm in many areas of cognitive science, regards the mind as a digital computer that stores and manipulates symbol sequences according to fixed rules of transformation. On this view, mental events, states, and processes are computation events, states, and processes, and there is nothing more to a cognitive process than what is captured in a computer program successfully modeling it. This perspective on minds and mental processes seems to entail—at least, it encourages—the claim that a computer running a program that models a human cognitive process is itself engaged in that cognitive process. Thus, a computer that successfully simulates college students constructing proofs in sentential logic is itself engaged in the activity of constructing logical proofs. As we saw earlier, machine functionalism holds that having a mind is being a physical Turing machine of appropriate complexity. The issue of "appropriateness" aside, it is clear that the route from machine functionalism to computationalism is pretty straight and short.

This view of computation and mind is what John Searle calls "strong AI," which he characterizes as follows:

> According to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states.[27]

Before we discuss Searle's intriguing argument against computationalism, we might wonder why anyone would conflate a simulation with the real thing being simulated. Computers are used to simulate many different things: the performance of a jet engine, the spread of rabies in wildlife, the progress of a hurricane, and on and on. But no one would confuse a computer simulating a jet engine with a jet engine, or a computer simulation of a tornado with a tornado. So why would anyone want to say that a computer simulation of a cognitive process is itself a cognitive process? Isn't this a simple confusion? The following answer is open to the computationalist: It is because a cognitive process itself is a computational process. This means that a computer simula-

tion of a cognitive process is a simulation of a computational process, and obviously a computational simulation of a computational process is to re-create that computational process. Thus, there is no confusion in the claim that a computer simulating a cognitive process is itself engaged in that cognitive process. But this response makes sense only if we have already accepted the claim that cognitive processes are computational processes—that is, the truth of computationalism. This is the view of mind that Searle's Chinese room argument is expressly designed to refute.

To prepare us for his argument, Searle describes a program developed by Roger Schank and his colleagues to model our ability to understand stories. It is part of this ability that we are able to answer questions about details of a story that are not explicitly stated. Searle gives two examples. In the first story you are told: "A man goes into a restaurant and orders a hamburger. When it arrives, it is burned to a crisp, and the man angrily leaves, without paying for the hamburger." If you are asked "Did the man eat the hamburger?" you would presumably say "No." The second story goes like this: "A man goes into a restaurant and orders a hamburger. When it arrives, he is very pleased, and when he leaves, he leaves a big tip for the waiter." If you are asked "Did the man eat the hamburger?" you would say "Yes, he did." Schank's program is designed to answer questions like these in appropriate ways when it is given the stories. To do this, it has in its memory information concerning restaurants and how people behave in restaurants, ordering dishes, tipping, and so on. For the sake of the argument, we may assume Schank's program works flawlessly—it works perfectly as a simulation of the human ability to understand stories. The claim of computationalism would then be that a computer running Schank's program *literally understands* stories, just as we do.[28]

To undermine this claim, Searle constructs an ingenious thought-experiment.[29] He invites us to imagine a room—the "Chinese room"—in which someone (say, Searle himself) who understands no Chinese is confined. There are two piles of Chinese texts in the room, one called "the script" (this corresponds to the background information about restaurants, etc., in Schank's program) and "the story" (corresponding to the story on which understanding is tested). Searle is provided with a set of rules stated in English (the "rule book," which corresponds to Schank's program) for systematically transforming strings of symbols, by consulting the script and the story, to yield further symbol strings. These symbol strings are made up of Chinese characters, and the transformation rules are purely *formal*, or *syntactic*, in that their applications depend solely on the shapes of the symbols involved, not their meanings. So you can apply these rules without knowing any Chinese (remember: the rules

are stated in English); all that is required is that you have the ability to recognize Chinese characters by their shapes. Searle becomes very adept at manipulating Chinese expressions in accordance with the rules given to him (we may suppose that Searle has memorized the whole rule book) so that every time a string of Chinese characters is sent in, Searle goes to work, consulting the two piles of Chinese texts in the room, and promptly sends out a string of Chinese characters. From the perspective of someone outside the room who knows Chinese, the input strings are questions in Chinese about the story and the output strings sent out by Searle are responses to these questions. The input-output relationships are what we would expect if a Chinese speaker, instead of Searle, were locked inside the room. And yet Searle does not know any Chinese and does not understand the story, and there is no understanding of Chinese going on anywhere inside the room. What goes on is only manipulation of symbols on the basis of their shapes, or "syntax," but real understanding involves "semantics," knowing what these symbols represent, or mean. Although Searle's behavior is input-output equivalent to that of a speaker of Chinese, Searle knows no Chinese and does not understand the story (remember: the story is in Chinese).

Now, replace Searle with a computer running Searle's rule book as its program. This changes nothing: Both Searle and the computer are syntax-driven machines manipulating strings of symbols according to their shapes. In general, what goes on inside a computer is exactly like what goes on in the Chinese room (with Searle in it): rule-governed manipulations of symbols based on their syntactic shapes. There is no more understanding of the story in the computer than there is in the Chinese room. The conclusion to be drawn, Searle argues, is that mentality is more than rule-governed syntactic manipulation of symbols and that there is no way to generate semantics—what the symbols mean or represent—from their syntax. This means that understanding and other intelligent mental states and activities cannot arise from mere syntactic processes. Anyway, that is Searle's Chinese room argument.

Searle's argument has elicited a large number of critical responses, and just what the argument succeeds in showing remains controversial. Although its intuitive appeal and power cannot be denied, we have to be cautious in assessing its significance. The appeal of Searle's example may be due, some have argued, to certain misleading assumptions tacitly made in the way he describes what is going on in the Chinese room. Searle himself describes, and tries to respond to, six "replies" to his argument. Some of the objections to Searle raise serious points, and the reader is urged to examine them and Searle's responses. These responses are often ingenious and thought-provoking; however, Searle tends to

appeal to the intuitions of his readers, and we could do more, it seems, to drive home his central point, namely the thesis that syntactical manipulations do not generate meanings, or anything that can be called an understanding of stories. Consider, then, the following reconstructed argument in behalf of Searle:

> (1) Let us begin by asking what exactly is the difference between, on one hand, Searle/the computer in the Chinese room and, on the other, a Chinese speaker. (We assume that the program being run is Schank's program modeling story understanding.)
>
> (2) To understand the two stories in Chinese about a man ordering a hamburger in a restaurant, you must know, among other things, that "煎牛肉饼" means hamburger.
>
> (3) The Chinese speaker knows this, but neither Searle nor the computer does. That is why the Chinese speaker understands the stories, but Searle and the computer do not.
>
> (4) No amount of syntactic manipulation of Chinese characters will enable someone to acquire the knowledge of what "煎牛肉饼" means.
>
> (5) Hence, computationalism is false; neither Searle nor the computer running Schank's program understands the stories.

The central idea is that knowledge of meaning, or semantic knowledge, involves *word-to-world* (or *language-to-world*) relationships, whereas syntax concerns only properties and relations *within* a language as a symbol system. To acquire meanings, you must break out of the symbol system into the real world. Pushing symbols around according to their shapes will not get you in touch with extralinguistic reality. To know that "煎牛肉饼" means hamburger, you have to know what hamburgers are, and you come by this knowledge only through real-life contact with hamburgers (eating a few will help), or through descriptions in terms of other things you know through your real-life experience. Syntactic symbol manipulation alone will not yield such knowledge; only real-world experience will.

To expect syntactic operations to generate knowledge of meaning is like trying to learn a new language, say Russian, by memorizing a Russian–Russian dictionary. Or consider this example: You memorize a Korean-Japanese dictionary, and it may be possible for you to translate any Korean sentence into

Japanese by following a set of formal rules (stated in English—you can memorize these rules, too, like Searle memorizing the rule book). (Think of the translation programs available on many websites.) But you do not understand a word of Korean, or a word of Japanese, though you have become a proficient translator between the two languages. To understand either language, you have to know how that language is hooked up with the things in the real world.

So far so good. We have to be cautious, though, about what our argument, if successful, shows. It only shows that the computer running Schank's program (sitting in the basement of some computer-science lab) has no understanding of the stories in Chinese. It does not show, as Searle thinks the Chinese room shows, that no computing machine, an electromechanical device running programs, can acquire semantic knowledge of the sort displayed in (2) above. What our argument suggests is that for a computing machine (or anything else) to acquire this kind of knowledge, it must be placed in the real world, interacting with its environment, acquiring information about its surroundings, and possibly interacting with other agents like itself. In short, it must be an android, not necessarily humanlike in appearance, but an agent and cognizer in real-life situations, like Commander Data in the television series *Star Trek: The Next Generation*. (How meanings arise is itself a big question in philosophy of mind and language; see chapter 8 on mental content.)

Searle, however, is of the opinion that meaning and understanding can arise only in biological brains,[30] a position he calls "biological naturalism." On this approach, neural states, those that underlie thoughts, will carry representational contents. However, it seems clear that there are no relevant differences between neural processes and computational processes that could tilt the case in favor of biology over electronics. The fact is that the same neurobiological causal processes will go on no matter what these neural states represent about the world or whether they represent anything at all. That is, neural processes are no more responsive to meaning and representational content than are electronic computational processes. Local physical-biological conditions in the brain, not the distal states of affairs represented by neural states, are what drive neural processes. If so, isn't Searle in the same boat as Turing and other computationalists?

The question, therefore, is not what drives computational processes or neural processes. In neither do meanings or contents play a causal role; it is only the syntactic shapes of symbolic representations and the intrinsic physicochemical properties of the brain states that drive the processes. The important question is how these representations and neural states acquire meanings and intentionality in the first place. This is where the contact with the real world enters the picture: What we can conclude with some confidence

at this point is that such contact is crucial if a system, whether a human person or a machine, is to gain capacities for speech, understanding, and other cognitive functions and activities.

For Further Reading

The classic source of machine functionalism is Hilary Putnam's "Psychological Predicates" (later reprinted as "The Nature of Mental States"). See also his "Robots: Machines or Artificially Created Life?" and "The Mental Life of Some Machines"; all three papers are reprinted in his *Mind, Language, and Reality: Philosophical Papers*, volume 2. The first of these is widely reprinted elsewhere, including *Philosophy of Mind: Classical and Contemporary Readings*, edited by David J. Chalmers, and *Philosophy of Mind: A Guide and Anthology*, edited by John Heil. Ned Block's "What Is Functionalism?" is a clear and concise introduction to functionalism. Putnam, the founder of functionalism, later renounced it; see his *Representation and Reality*, chapters 5 and 6.

For a teleological approach to functionalism, see William G. Lycan, *Consciousness*, chapter 4. For a general biological-evolutionary perspective on mentality, Ruth G. Millikan's *Language, Thought, and Other Biological Categories* is an important source.

For issues involving the Turing test and the Chinese room argument, see Alan M. Turing, "Computing Machinery and Intelligence"; John R. Searle, "Minds, Brains, and Programs"; and Ned Block, "The Mind as Software in the Brain." These articles are reprinted in Heil's *Philosophy of Mind*. Also recommended are Block, "Psychologism and Behaviorism," and Daniel C. Dennett, *Consciousness Explained*, chapter 14. Entries on "Turing Test" and "Chinese Room Argument" in the *Stanford Encyclopedia of Philosophy* are useful resources.

For criticisms of machine functionalism (and functionalism in general), see Ned Block, "Troubles with Functionalism," and John R. Searle, *The Rediscovery of the Mind*.

Notes

1. Later given a new title "The Nature of Mental States."

2. Donald Davidson's argument for mental anomalism (chapter 7) also played a part in the decline of reductionism. See Davidson's "Mental Events."

3. At least some of them, for it could be argued that certain psychological states can be had only by materially embodied subjects—for example, feelings of hunger and thirst, bodily sensations like pain and itch, and sexual desire.

4. Unless, that is, the very idea of an immaterial mental being turns out to be incoherent.

5. The terms "realize," "realization," and "realizer" are explained explicitly in a later section. In the meantime, you will not go far astray if you read "P realizes M" as "P is a neural substrate, or base, of M."

6. This principle entails mind-body supervenience, which we characterized as minimal physicalism in chapter 1. Further, it arguably entails the thesis of ontological physicalism, as stated in that chapter.

7. See Ronald Melzack, *The Puzzle of Pain*, pp. 15–16.

8. Some have argued that this function-versus-mechanism dichotomy is pervasive at all levels, not restricted to the mental-physical case; see, for example, William G. Lycan, *Consciousness*.

9. As I take it, something like this is the point of Karl Lashley's principle of "equipotentiality"; see his *Brain Mechanisms and Intelligence*, p. 25.

10. To borrow Ned Block's question in "What Is Functionalism?" pp. 178–179.

11. As we shall see in connection with machine functionalism, there is another sense of "function," the mathematical sense, involved in "functionalism."

12. Strictly speaking, it is more accurate to say that having *the capacity to sense pain* is being equipped with a tissue-damage detector, and that pain, as an occurrence, is the activation of such a detector.

13. The distinction between "real" and "nominal" essence goes back to John Locke. A full explanation of these notions cannot be provided here. See Locke, *An Essay on Human Understanding*, Book III, chapters iii and vi. For helpful discussion see Nicholas Jolley, *Locke: His Philosophical Thought*, chapters 4 and 8.

14. When do two mousetraps count as instances of the same realizer and when do they count as instances of different realizers? What about pains and their realizers? These are significant questions. For helpful discussion see Lawrence Shapiro, *The Mind Incarnate*.

15. See, for example, B. F. Skinner, "Selections from *Science and Human Behavior*."

16. Machine functionalism in terms of Turing machines developed in sections below can deal with this problem as well; however, the Ramsey-Lewis method presented in chapter 6 is more intuitive and perspicacious.

17. A treatment of the mathematical theory of computability in terms of Turing machines can be found in Martin Davis, *Computability and Unsolvability*, and in George S. Boolos, John P. Burgess, and Richard C. Jeffrey, *Computability and Logic*.

18. Strictly speaking, this was a proposal, called the Church-Turing Thesis, rather than a discovery. It turned out that various proposed notions of "effective"

or "mechanical" calculability, including computability by a Turing machine, turned out to be mathematically equivalent, defining the same class of functions. The thesis was the proposal that these notions of effective calculable functions be taken as equivalent ways of defining "computable" functions. For details see the entry "Church-Turing Thesis" in the *Stanford Encyclopedia of Philosophy*.

19. See, for example, Hilary Putnam, "Psychological Predicates."

20. For a statement and defense of a position of this kind, see Bas Van Fraassen, *The Scientific Image*.

21. Is there, for any given psychological subject, a *unique* Turing machine that is a machine description (relative to a specification of input and output conditions), or can there be (perhaps there always must be) multiple, nontrivially different machine descriptions? Does realism about psychology require that there be a unique one?

22. Alan M. Turing, "Computing Machinery and Intelligence."

23. It is probably more reasonable to restrict the claim to cognitive mentality, leaving out things like sensations and emotions.

24. To do well on a real-life Turing test, the computers will need to have a real-time processing speed, in addition to delivering the "right" output (answers) for the given input (questions).

25. For an elaboration and discussion of this point, see Ned Block, "Psychologism and Behaviorism."

26. Daniel C. Dennett, *Consciousness Explained*, pp. 435–440.

27. John R. Searle, "Minds, Brains, and Programs," p. 235 in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

28. Here we are setting aside an important question discussed earlier, namely that of psychological reality. Is Schank's program merely input-output equivalent to human understanding of stories, or does it in some relevant sense mirror the actual cognitive processes involved in human story understanding?

29. John R. Searle, "Minds, Brains, and Programs."

30. Or, says Searle, structures (even computers) that have the same causal powers as brains. My brain, in virtue of its weight, has the causal power of breaking eggs when dropped on them. But surely having this causal power cannot be relevant to mentality. So just what causal powers of a brain must a thing have in order to enjoy a mental life? Obviously, the brain's powers to generate and sustain a mental life! As it stands, therefore, Searle's apparent concession on the biological basis of mentality isn't very helpful.

CHAPTER 6

Mind as a Causal System

Causal-Theoretical Functionalism

In the preceding chapter, we discussed the functionalist attempt to use Turing machines to explicate the nature of mentality and its relationship to the physical. Here we examine another formulation of functionalism, in terms of "causal role." Central to any version of functionalism is the idea that a mental state can be characterized in terms of the input-output relations it causally mediates, where the inputs and outputs may include other mental states as well as sensory stimuli and physical behaviors. Mental phenomena are conceived as nodes in a complex causal network that engages in causal transactions with the outside world at its peripheries, by receiving sensory inputs and emitting behavior outputs.

What, according to functionalism, distinguishes one mental kind (say, pain) from another (say, itch) is the distinctive input-output relationship associated with each kind. Causal-theoretical functionalism conceives of this input-output relationship as a causal relation, one that is mediated by mental states. Different mental states are different because they are implicated in different input-output causal relationships. Pain differs from itch in that each has its own distinctive causal role: Pains typically are caused by tissue damage and cause winces, groans, and escape behavior; in contrast, itches typically are caused by skin irritation and cause scratching. But tissue damage causes pain only if certain other conditions are present, some of which are mental in their own right; not only must you have a properly functioning nervous system, but you must also be normally alert and not engrossed in another task. Moreover, among the typical effects of pain are further mental states, such as a feeling of distress and a desire to be relieved of it. But this seems to involve

us in a regress or circularity: To explain what a given mental state is, we need to refer to other mental states, and explaining these can only be expected to require reference to further mental states, and so on—a process that can go on in an unending regress or loop back in a circle. Circularity threatens to arise at a more general level as well, in the functionalist conception of mentality itself: To be a mental state is to be an internal state serving as a causal intermediary between sensory inputs and *mental states* as causes, on the one hand, and behaviors and other *mental states* as effects, on the other. Viewed as a definition of what it is to be a mental state, this is obviously circular. To circumvent the threatened circularity, machine functionalism exploits the concept of a Turing machine in characterizing mentality. To achieve the same end, causal-theoretical functionalism exploits the entire network of causal relations involving all psychological states—in effect, a comprehensive psychological theory—to anchor the physical-behavioral definitions of individual mental properties.[1]

The Ramsey-Lewis Method

Consider the following toy "pain theory":

> (T) For any *x*, if *x suffers tissue damage* and **is normally alert**, *x* **is in pain**; if *x is awake*, *x* tends to be **normally alert**; if *x* **is in pain**, *x winces* and *groans* and **goes into a state of distress**; and if *x* **is not normally alert** or *x* **is in a state of distress**, *x tends to make more typing errors*.

We assume that the statements constituting T describe lawful regularities (or causal relations). The italicized expressions are nonmental predicates designating observable physical, biological, and behavioral properties; the expressions in boldface are psychological predicates designating mental properties. T is, of course, much less than what we know about pain and its relationship to other events and states, but let us assume that T encapsulates what is important about our knowledge of pain. Issues about the kind of "theory" T must be if T is to serve as a basis of functional definitions of mental expressions will be taken up in a later section. Here T is only an example to illustrate the formal technique originally due to Frank P. Ramsey, a British mathematician-philosopher in the early twentieth century, and later adapted by David Lewis for formulating functional definitions of mental kinds.[2]

We first "Ramseify" T by "existentially generalizing" over each psychological predicate occurring in it, which yields this:

(T_R) There exist states M_1, M_2, and M_3 such that for any *x*, if *x suffers tissue damage* and is in M_1, *x* is in M_2; if *x is awake*, *x* tends to be in M_1; if *x* is in M_2, *x winces* and *groans* and goes into M_3; and if *x* is either not in M_1 or is in M_3, *x tends to make more typing errors*.

The main thing to notice about T_R vis-à-vis T is that instead of referring (as T does) to specific mental states, T_R speaks only of *there being some states or other*, M_1, M_2, and M_3, which are related to each other and to observable physical-behavioral states in the way specified by T. Evidently, T logically implies T_R (essentially in the manner in which "*x* is in pain" logically implies "There is some state M such that *x* is in M"). Note that in contrast to T, its Ramseification T_R contains no psychological expressions but only physical-behavioral expressions such as "suffers tissue damage," "winces," and so on. Terms like "M_1," "M_2," and "M_3" are called predicate variables (they are like the *x*s and *y*s in mathematics, though these are usually used as "individual" variables)—they are "topic-neutral" logical terms, neither physical nor psychological. Expressions like "is normally alert" and "is in pain" are predicate constants, that is, actual predicates.

Ramsey, who invented the procedure now called "Ramseification," showed that although T_R is weaker than T (since it is implied by, but does not imply, T), T_R is just as powerful as T as far as physical-behavioral prediction goes; the two theories make exactly the same deductive connections between nonpsychological statements.[3] For example, both theories entail that if someone is awake and suffers tissue damage, she will wince, and that if she does not groan, either she has not suffered tissue damage or she is not awake. Since T_R is free of psychological expressions, it can serve as a basis for defining psychological expressions without circularity.

To make our sample definitions manageable, we abbreviate T_R as "$\exists M_1, M_2, M_3[T(M_1, M_2, M_3)]$." (The symbol $\exists$, called the "existential quantifier," is read: "there exist.") Consider, then:[4]

x is in pain = $_{def}$ $\exists M_1, M_2, M_3[T(M_1, M_2, M_3)$ and *x* is in $M_2]$

Note that "M_2" is the predicate variable that replaced "is in pain" in T. Similarly, we can define "is alert" and "is in distress" (although our little theory T was made up mainly to give us a reasonable definition of "pain"):

x is normally alert = $_{def}$ $\exists M_1, M_2, M_3$ [T(M_1, M_2, M_3) and *x* is in M_1]

x is in distress = $_{def}$ $\exists M_1, M_2, M_3$ [T(M_1, M_2, M_3) and *x* is in M_3]

Let us see what these definitions say. Consider the definition of "being in pain": It says that you are in pain just in case there are certain states, M_1, M_2, and M_3, that are related among themselves and with such physical-behavioral states as tissue damage, wincing and groaning, and typing performance as specified in T_R *and* you are in M_2. It is clear that this definition gives us a concept of pain in terms of its causal-nomological relations and that among its causes and effects are other "mental" states (although these are not specified as such but referred to only as "some" states of the psychological subject) as well as physical and behavioral events and states. Notice also that there is a sense in which the three mental concepts are interdefined but without circularity; each of the defined expressions is completely eliminable by its definiens (the right-hand side of the definition), which is completely free of psychological expressions. Whether or not these definitions are adequate in all respects, it is evident that the circularity problem has been solved.

So the trick is to define psychological concepts holistically en masse. Our T is a fragment of a theory, something made up to show how the method works; to generate more realistic functional definitions of psychological concepts by the Ramsey-Lewis method, we need a comprehensive underlying psychological theory encompassing many more psychological kinds and richer and more complex causal-nomological relationships to inputs and outputs. Such a theory will be analogous to a Turing machine that models a full psychology, and the resemblance of the present method with the approach of machine functionalism should be clear, at least in broad outlines. In fact, we can think of the Turing machine approach as a special case of the Ramsey-Lewis method in which the psychological theory is presented in the form of a Turing machine table with the internal machine states, the *q*s, corresponding to the predicate variables, the Ms. We discuss the relationship between the two approaches in more detail later.

Choosing an Underlying Psychology

So what should the underlying psychological theory T be like if it is to yield, by the Ramsey-Lewis technique, adequate functional definitions of psychological properties? If we are to recover a psychological property from T_R by the Ramsey-Lewis method, the property must appear in T to begin with. So T must refer to all psychological properties. Moreover, T must carry enough information about each psychological property—about how it is nomologically connected with input conditions, behavior outputs, and other psychological

properties—to circumscribe it closely enough to identify it. Given this, there are two major possibilities to consider.

We might, with Lewis, consider using the platitudes of our shared *commonsense psychology* as the underlying theory. The statements making up our "pain theory" T are examples of such platitudes, and there are countless others about, for instance, what makes people angry and how angry people behave, how wants and beliefs combine to generate further wants, how perceptions cause beliefs and memories, and how beliefs lead to further beliefs. Few people are able to articulate these principles of "folk psychology," but most mature people use them constantly in attributing mental states to people, making predictions about how people will behave, and understanding why people do what they do. We know these psychological regularities "tacitly," perhaps in much the way we "know" the grammar of the language we speak—without being able to state any explicit rules. Without a suitably internalized commonsense psychology in this sense, we would hardly be able to manage our daily transactions with other people and enjoy the kind of communal life that we take for granted.[5] It is important that the vernacular psychology that serves as the underlying theory for functional definitions consists of *commonly known* generalizations. This is essential if we are to ensure that functional definitions yield the psychological concepts that all of us share. It is the shared funds of vernacular psychological knowledge that collectively define our commonsense mental concepts; there is no other conceivable source from which our mental concepts could magically spring. Functionalism that takes these psychological platitudes as a basis for functional definitions of psychological terms is sometimes called "analytical functionalism." The thought is that these well-known psychological generalizations are virtually "analytic" truths—truths that are evident to speakers who understand the meanings of the psychological expressions involved.

We must remember that commonsense psychology is, well, only commonsensical: It may be incomplete and partial, and contain serious errors, or even inconsistencies. If mental concepts are to be defined in terms of causal-nomological relations, shouldn't we use our best theory about how mental events and states are involved in causal-nomological relations, among themselves and with physical and behavioral events and processes? *Scientific psychology*, including cognitive science, after all, is in the business of investigating these regularities, and the best scientific psychology we can muster *is* the best overall theory about the causal-nomological facts of mental events and states. The form of functionalism that favors empirical scientific theory as the Ramseification base is sometimes called "psycho-functionalism."

There are problems and difficulties with each of these choices. Let us first note one important fact: If the underlying theory T is false, we cannot count on any mental concepts defined on its basis to apply to anything—as logicians will say, these concepts will have empty, or null, extensions.[6] For if T is false, its Ramseification, T_R, may also be false; in particular, if T has false nonmental consequences (for example, T makes wrong behavioral predictions), T_R will be false as well. (Recall that T and T_R have the same physical-behavioral content.) If T_R is false, every concept defined on its basis by the Ramsey-Lewis method will be vacuous—that is, it will not apply to anything. This is easy to see for our sample "pain theory" T. Suppose this theory is false—in particular, suppose that what T says about the state of distress is false and that in fact there is no state that is related, in the way specified by T for distress, with the other internal states and inputs and outputs. This makes our sample T_R false as well, since there is nothing that can fill in for M_3. This would mean that "pain" as defined on the basis of T_R cannot be true of anything: Nothing satisfies the defining condition of "pain." The same goes for "normally alert" and "the state of distress." So if T, the underlying theory, is false, all mental concepts defined on its basis by the Ramsey-Lewis method will turn out to have the same extension, namely, the null extension!

This means that we had better make sure that the underlying theory is true. If our T is to yield our psychological concepts all at once, it is going to be a long conjunction of myriad psychological generalizations, and even a single false component will make the whole conjunction false. So we must face these questions: What is going to be included in our T, and how certain can we be that T is true? Consider the case of scientific psychology: It is surely going to be a difficult, probably impossible, task to decide what parts of current scientific psychology are well enough established to be considered uncontroversially true. Psychology has been flourishing as a science for many decades now, but it is comparatively young as a science, with its methodological foundations still in dispute, and it is fair to say that it has yet to produce a robust enough common core of generally accepted laws and theories. In this respect, psychology has a long way to go before it reaches the status of, say, physics, chemistry, or even biology.

These reflections lead to the following thought: On the Ramsey-Lewis method of defining psychological concepts, every dispute about the underlying theory T is going to be a dispute about psychological concepts. This creates a seemingly paradoxical situation: If two psychologists should disagree about some psychological generalization that is part of theory T, which we should expect to be a common occurrence, this would mean that they are us-

ing different sets of psychological concepts. But this seems to imply that they cannot really disagree, since the very possibility of disagreement presupposes that the same concepts are shared. How could I accept and you reject a given proposition unless we shared the concepts in terms of which the proposition is formulated?

Perhaps things are not as bleak as they seem. For example, there is probably no need to invoke a total psychology as a base for functional definitions of mental terms; relatively independent parts of psychology and cognitive science, like theory of vision, theory of motivation, decision and action, theory of language acquisition, and so on, can each serve as a basis of Ramseification. Also we can consider degrees of similarity between concepts, and it may be possible for two speakers to understand each other well enough in a given situation, without sharing an exactly identical set of concepts; sharing similar concepts may be good enough for the purposes at hand.

Consider again the option of using commonsense psychology to anchor psychological concepts. Can we be sure that all of our psychological platitudes, or even any of them, are true—that is, that they hold up as systematic scientific psychology makes progress? Some have even argued that advances in scientific psychology have already shown commonsense psychology to be massively erroneous and that, considered as a theory, it must be abandoned.[7] Consider the generalization, used as part of our pain theory, that tissue damage causes pain in a normally alert person. It is clear that there are many exceptions to this regularity: A normally alert person who is totally absorbed in another task may not feel pain when she suffers minor tissue damage. Massive tissue damage may cause a person to go into a coma. And what is to count as "normally alert" in any case? Alert enough to experience pain when one is hurt? The platitudes of commonsense psychology may serve us competently enough in our daily life in anticipating behaviors of our fellow humans and making sense of them. But are we prepared to say that they are literally true? One way to alleviate these worries is to point out that we should think of our folk-psychological generalizations as hedged by generous escape clauses ("all other things being equal," "under normal conditions," "in the absence of interfering forces," and so on). Whether such weakened, noncommittal generalizations can introduce sufficiently restrictive constraints to yield well-defined psychological concepts is something to think about.

In one respect, though, commonsense psychology seems to have an advantage over scientific psychology: its apparently greater stability. Theories and concepts of systematic psychology come and go; given what we know about the rise and fall of scientific theories, especially in the social and human sciences, it

is reasonable to expect that most of what we now consider our best theories in psychology will be abandoned and replaced, or seriously revised, sooner or later—probably sooner rather than later. The rough regularities codified in commonsense psychology appear considerably more stable (perhaps because they are rough); can we really imagine giving up the virtual truism that a person's desire for something and her belief that doing a certain thing will secure it tends to cause her to do it? This basic principle, which links belief and desire to action, is a central principle of commonsense psychology that makes it possible to understand why people do what they do. It seems reasonable to think that the principle was as central to the way the ancient Greeks or Chinese made sense of themselves and their fellows as it is to our own folk-psychological explanatory practices. Our shared folk-psychological heritage is what enables us to understand, and empathize with, the actions and emotions of the characters depicted in Greek tragedies and historical Chinese fiction. Indeed, if there were a culture, past or present, for whose members the central principles of our folk psychology, such as the one that relates belief and desire to action, did not hold true, its institutions and practices would hardly be intelligible to us, and its language might not even be translatable into our own. The source and nature of this relative permanence and commonality of folk-psychological platitudes are in need of explanation, but it seems plausible that folk psychology enjoys a degree of stability and universality that eludes scientific psychology.

We should note, though, that vernacular psychology and scientific psychology need not necessarily be thought to be in competition with each other. We could say that vernacular psychology is the appropriate underlying theory for the functional definition of vernacular psychological concepts, while scientific psychology is the appropriate one for scientific psychological concepts. If we believe, however, that scientific psychology shows, or has shown, vernacular psychology to be seriously flawed (for example, showing that many of its central generalizations are in fact false),[8] we would have to reject the utility of the concepts generated from it by the Ramsey-Lewis method, for as we saw, these concepts would then apply to nothing.

There is one final point about our sample functionalist definitions: They can accommodate the phenomenon of multiple realization of mental states. This is easily seen. Suppose that the original pain theory, T, is true of both humans and Martians, whose physiology, let us assume, is very different from ours (it is inorganic, say). Then T_R, too, would be true for both humans and Martians: It is only that the triple of physical-biological states $<H_1, H_2, H_3>$, which realizes the three mental states <pain, normal alertness, distress> and therefore satisfies T_R for humans, is different from the triple of physical states

$<I_1, I_2, I_3>$, which realizes the mental triple in Martians. But in either case there exists a triple of states that are connected in the specified ways, as T_R demands. So when you are in H_2, you are in pain, and when Mork the Martian is in I_2, he is in pain, since each of you satisfies the functionalist definition of pain as stated.

Functionalism as Physicalism: Psychological Reality

Let us return to scientific psychology as the underlying theory to be Ramseified. As we noted, we want this theory to be a true theory. Now, there is another question about the truth of psychological theories that we need to attend to. Let us assume that psychological theories posit internal states to systematize correlations between sensory inputs and behavioral outputs. These internal states are the putative psychological states of the organism. Suppose now that each of two theories, T_1 and T_2, gives a correct systematization of inputs and outputs for a psychological subject S, but that each posits a different set of internal states. That is, T_1 and T_2 are both *behaviorally adequate* psychologies for S, but each attributes to S a different internal causal structure connecting S's inputs to its outputs. Is there some further fact about these theories, or about S, that determines which (if any) is the correct psychology of S? As a basis for Ramseified functional definitions of mental states, we presumably must choose the correct psychology *if* there is a correct one.

If psychology is a truly autonomous special science, under no methodological, theoretical, or metaphysical constraints from any other science, we would have to say that the only ground for preferring one or the other of two behaviorally adequate theories consists in broad formal considerations of notational simplicity, ease of deriving predictions, and the like. There could be no further fact-based grounds favoring one theory over the other. As you will recall, behaviorally adequate psychologies for subject S are analogous to Turing machines that are "behavioral descriptions" of S (see chapter 5). You will also recall that according to machine functionalism, not every behavioral description of S is a correct psychology of S and that a correct psychology is one that is a machine description of S—namely, a Turing machine that is physically realized by S. This means that there are internal physical states of S that realize the internal machine states of the Turing machine in question—that is, there are in S "real" internal physical states that are (causally) related to each other and to sensory inputs and behavioral outputs as specified by the machine table of the Turing machine. It is the requirement of physical realization

that answers the question of the psychological reality of Turing machines purporting to specify the psychologies of a subject.

Unlike machine functionalism, causal-theoretical functionalism, formulated on the Ramsey-Lewis model, does not as yet have a physical requirement built into it. According to machine functionalism as formulated in the preceding chapter, for subject S to be in any mental state, S must be a *physical realization* of an appropriate Turing machine; in contrast, causal-theoretical functionalism as developed thus far in this chapter requires only that there be "internal states" of S that are connected among themselves and to inputs and outputs as specified by S's psychology, without saying anything about the nature of these internal states. What we saw in connection with machine functionalism was that it is the further physical requirement—to the effect that the states of S that realize the machine's internal states be physical states—that makes it possible to pick out S's correct psychology. In the same way, the only way to settle the issue of psychological reality between behaviorally adequate psychologies is to explicitly introduce a similar physicalist requirement, perhaps something like this:

> (P) The states that the Ramseified psychological theory, T_R, affirms to exist are physical-neural states; that is, the variables M_1, M_2, . . . of T_R and in the definitions of specific mental states (see our sample definitions of "pain," and so on) range over physical-neural states of the subjects of psychological theory T.

A functionalist who accepts (P) may be called a physicalist functionalist. Unless some physical constraints, represented by (P), are introduced, there seems to be no way of discriminating between behaviorally adequate psychologies. Conversely, the apparent fact that we do not think all behaviorally adequate psychologies are "correct" or "true" signifies our commitment to the reality of the internal, theoretical states posited by our psychologies, and the only way this psychological realism is cashed out is to regard these states as internal *physical* states of the organism involved. This is equivalent in substance to the thesis of realization physicalism discussed in the preceding chapter—the thesis that all psychological states, if realized, must be physically realized.

This appears to reflect the actual research strategies in psychology and cognitive science and the methodological assumptions that undergird them: The correct psychological theory must, in addition to being behaviorally adequate, have "physical reality" in the sense that the psychological capacities, dispositions, and mechanisms it posits have a physical (presumably neurobiological)

basis. The psychology that gives the most elegant and simplest systematization of human behavior may not be the true psychology, any more than the simplest artificial intelligence program (or Turing machine) that accomplishes a certain intelligent task (proving logic theorems, face recognition, or whatever) accurately reflects the way we humans perform it. The psychological theory that is formally the most elegant may not describe the way humans (or other organisms or systems under consideration) actually process their sensory inputs and produce behavioral outputs. There is no reason, either a priori or empirical, to believe that the mechanism that underlies our psychology, something that has evolved over many millions of years in the midst of myriad unpredictable natural forces, must be in accord with our notion of what is simple and elegant in a scientific theory. The psychological capacities and mechanisms posited by a true psychological theory must be real, and the only reality to which we can appeal in this context seems to be physical reality. These considerations, quite apart from the arguments pro and con concerning the physical reducibility of psychology, cast serious doubts on the claim that psychology is an autonomous science not answerable to lower-level physical-biological sciences.

The antiphysicalist might argue that psychological capacities and mechanisms have their own separate, nonphysical reality. But it is difficult to imagine what they could be when dissociated from their physical underpinnings; could they be some ghostly mechanisms in Cartesian mental substances? That may be a logically possible position, but hardly a plausible one, philosophically or scientifically (see chapter 2). It isn't for nothing that physicalism is the default position in contemporary philosophy of mind and psychology.

Objections and Difficulties

In this section, we review several points that are often thought to present major obstacles to the functionalist program. Some of the problematic features of machine functionalism discussed in the preceding chapter apply, mutatis mutandis, to causal-role functionalism, and these will not be taken up again here.

Qualia Inversion

Consider the question: What do all instances of pain have in common in virtue of which they are pains? You will recognize the functionalist answer: their characteristic causal role—their typical causes (tissue damage, trauma) and effects (pain behavior). But isn't there a more obvious answer? What all instances of pain have in common in virtue of which they are all cases of pain

is that they *hurt*. Pains hurt, itches itch, tickles tickle. Is there anything more obvious than that?

Sensations have characteristic *qualitative* features; these are called "phenomenal" or "phenomenological" or "sensory" qualities; "qualia" ("quale" for singular) is now the standard term. Seeing a ripe tomato has a certain distinctive visual quality that is unmistakably different from the visual quality involved in seeing a mound of spinach leaves. We are familiar with the smells of roses and ammonia; we can tell the sound of a drum from that of a gong; the feel of a cool, smooth granite countertop as we run our fingers over it is distinctively different from the feel of sandpaper. Our waking life is a continuous feast of qualia—colors, smells, sounds, and all the rest. When we temporarily lose our ability to taste or smell properly because of a bad cold, eating a favorite food can be like chewing cardboard and we are made acutely aware of what is missing from our experience.

By identifying sensory events with causal roles mediating input and output, however, functionalism appears to miss their qualitative aspects altogether. For it seems quite possible that causal roles and phenomenal qualities come apart, and the possibility of "qualia inversion" seems to prove it. It would seem that the following situation is perfectly coherent to imagine: When you look at a ripe tomato, your color experience is like my color experience when I look at a bunch of spinach, and vice versa. That is, your experience of red might be qualitatively like my experience of green, and your experience of green is like my experience of red. These differences need not show up in any observable behavioral differences: We both say "red" when we are shown ripe tomatoes, and we both describe the color of spinach as "green"; we are equally good at picking tomatoes out of mounds of lettuce leaves; and when we drive, we cope equally well with the traffic lights. In fact, we can coherently imagine that your color spectrum is systematically inverted with respect to mine, without this being manifested in any behavioral differences. Moreover, it seems possible to think of a system, like an electromechanical robot, that is functionally—that is, in terms of inputs and outputs—equivalent to us but to which we have no good reason to attribute any qualitative experiences (again, think of Commander Data; this is called the "absent qualia" problem).[9] If inverted qualia, or absent qualia, are possible in functionally equivalent systems, qualia cannot be captured by functional definitions, and functionalism cannot be an account of all psychological states and properties. This is the qualia argument against functionalism.

Can the functionalist offer the following reply? On the functionalist account, mental states are realized by the internal physical states of the psychological subject; so for humans, the experience of red, as a mental state, is

realized by a specific neural state. This means that you and I cannot differ in respect of the qualia we experience as long as we are in the same neural state; given that both you and I are in the same neural state, something that is in principle ascertainable by observation, either both of us experience red or neither does.

But this reply falls short for two reasons. First, even if it is correct as far as it goes, it does not address the qualia issue for physically different systems (say, you and the Martian) that realize the same psychology. Nothing it says makes qualia inversion impossible for you and the Martian; nor does it rule out the possibility that qualia are absent from the Martian experience. Second, the reply assumes that qualia supervene on the physical-neural states that realize them, but this supervenience assumption is part of what is at issue. However, the issue about qualia supervenience concerns the broader issues about physicalism; it is not specifically a problem with functionalism.

This issue concerning qualia has been controversial, with some philosophers doubting the coherence of the very idea of inverted or absent qualia.[10] We return to the issue of qualia in connection with the more general questions about consciousness (chapters 9 and 10).

The Cross-Wired Brain

Let us consider the following very simple, idealized model of how pain and itch mechanisms work: Each of us has a "pain box" and an "itch box" in our brains. We can think of the pain box as consisting of a bundle of neural fibers somewhere in the brain that gets activated when we experience pain, and similarly for the itch box. When pain receptors in our tissues are stimulated, they send neural signals up the pain input channel to the pain box, which then gets activated and sends signals down its output channel to our motor systems to cause appropriate pain behavior (winces and groans). The itch mechanism works similarly: When a mosquito bites you, your itch receptors send electrochemical signals up the itch input channel to your itch box, and so on, finally culminating in your itch behavior (scratching).

Suppose that a mad neurosurgeon rewires your brain by crisscrossing both the input and output channels of your pain and itch centers. That is, the signals from your pain receptors now go to your (former) itch box and the signals from this box now trigger your motor system to emit winces and groans; similarly, the signals from your itch receptors are now routed to your (former) pain box, which sends its signals to the motor system, causing scratching behavior. And suppose that I escape the mad neurosurgeon's attention. It is clear that

even though your brain is cross-wired with respect to mine, we both realize the same functional psychology: We both scratch when bitten by mosquitoes, and wince and groan when our fingers are burned. From the functionalist point of view, we instantiate the same pain-itch psychology.

Suppose that we both step barefoot on an upright thumbtack; both of us give out a sharp shriek of pain and hobble to the nearest chair. I am in pain. But what about you? The functionalist says that you, with the cross-wired brain, are in pain also. What makes a neural mechanism inside the brain a pain box is exactly the fact that it receives input from pain receptors and sends output to cause pain behavior. With the cross-wiring of your brain, your former itch box has now become your pain box, and when it is activated, you are in pain. At least that is what the functionalist conception of pain implies. But is this an acceptable consequence?

This is a version of the inverted qualia problem: Here the qualia that are inverted are pain and itch (or the painfulness of pains and the itchiness of itches), where the supposed inversion is made to happen through anatomical intervention. Many will feel a strong pull toward the thought that if your brain has been cross-wired as described, what you experience when you step on an upright thumbtack is an itch, not a pain, in spite of the fact that the input-output relation that you exhibit is one that is appropriate for pain. The appeal of this hypothesis is, at bottom, the appeal of the psychoneural identity theory of mentality. Most of us have a strong, if not overwhelming, inclination to think that types of conscious experience, such as pain and itch, supervene on the *local* states and processes of the brain no matter how they are hooked up with the rest of the body or the external world, and that the qualitative character of our mental states is conceptually and causally independent of their causal roles in relation to sensory inputs and behavioral outputs. Such an assumption is implicit, for example, in the popular philosophical thought-experiment with "the brain in a vat," in which a brain detached from a human body is kept alive in a vat of liquid and maintained in a normal state of consciousness by being fed electric signals generated by a supercomputer. The qualia we experience are causally dependent on the inputs: As our neural system is presently wired, cuts and pinpricks cause pains, not itches. But this is a contingent fact about our neural circuitry: It seems perfectly conceivable (even technically feasible at some point in the future) to reroute the causal chains involved so that cuts and pinpricks cause itches, not pains, and skin irritations cause pains, not itches, without disturbing the overall functional organization of our behavior.

Functional Properties, Disjunctive Properties, and Causal Powers

The functionalist claim is often expressed by assertions like "Mental states are causal roles" and "Mental properties (kinds) are functional properties (kinds)." We should get clear about the logic and ontology of such claims. The concept of a functional property and related concepts were introduced in the preceding chapter, but let us briefly review them before we go on with some difficulties and puzzles for functionalism. Begin with the example of pain: For something, S, to be in pain (that is, for S to have, or instantiate, the property of being in pain) is, according to functionalism, for S to be in some state (or to instantiate some property) with causal connections to appropriate inputs (for example, tissue damage, trauma) and outputs (pain behavior). For simplicity, let us talk uniformly in terms of *properties* rather than *states*. We may then say: The property of being in pain is the property of having some property with a certain causal specification, in terms of its causal relations to certain inputs and outputs. Thus, in general, we have the following canonical expression for all mental properties:

> Mental property M is the property of having a property with causal specification H.

As a rule, the functionalist believes in the multiple realizability of mental properties: For every mental property M, there will in general be multiple properties, Q_1, Q_2, . . . , each meeting the causal specification H, and an object will count as instantiating M just in case it instantiates one or another of these Qs. As you may recall, a property defined the way M is defined is often called a "second-order" property; in contrast, the Qs, their realizers, are "first-order" properties. (No special meaning needs to be attached to the terms "first-order" and "second-order"; these are relative terms—the Qs might themselves be second-order relative to another set of properties.) If M is pain, then its first-order realizers are neural properties, at least for organisms, and we expect them to vary across various pain-capable biological species.

This construal of mental properties as second-order properties seems to create some puzzles. If M is the property of having some property meeting specification H, where Q_1, Q_2, . . . , are the properties satisfying H—that is, the Qs are the realizers of M—it would seem to follow that M is identical with the *disjunctive* property of having Q_1 or Q_2 or . . . Isn't it evident that to have M just

is to have either Q_1 or Q_2 or . . . ? (For example, red, green, and blue are primary colors. Suppose something has a primary color; doesn't that amount simply to having red or green or blue?) Most philosophers who believe in the multiple realizability of mental properties deny that mental properties are disjunctive properties—disjunctions of their realizers—for the reason that the first-order realizing properties are extremely diverse and heterogeneous, so much so that their disjunction cannot be considered a well-behaved property with the kind of systematic unity required for propertyhood. As you may recall, the rejection of such disjunctions as legitimate properties was at the heart of the multiple realization argument against psychoneural-type physicalism. Functionalists have often touted the phenomenon of multiple realization as a basis for the claim that the properties studied by cognitive science are formal and abstract—abstracted from the material compositional details of the cognitive systems. What our considerations appear to show is that cognitive science properties so conceived threaten to turn out to be heterogeneous disjunctions of properties after all. And these disjunctions seem not to be suitable as nomological properties—properties in terms of which laws and causal explanations can be formulated. If this is right, it would disqualify mental properties, construed as second-order properties, as serious scientific properties.

But the functionalist may stand her ground, refusing to identify second-order properties with the disjunctions of their realizers, and she may reject disjunctive properties in general as bona-fide properties, on the ground that from the fact that both P and Q are properties, it does not follow that there is a disjunctive property, that of having P or Q. From the fact that being round and being green are properties, it does not follow, some have argued, that there is such a property as being round or green; some things that have this "property" (say, a red round table and a green square doormat) have nothing in common in virtue of having it. However, we need not embroil ourselves in this dispute about disjunctive properties, for the issue here is independent of the question about disjunctive properties.

For there is another line of argument, based on broad causal considerations, that seems to lead to the same conclusion. It is a widely accepted assumption, or at least a desideratum, that mental properties have causal powers: Instantiating a mental property can, and does, cause other events to occur (that is, cause other properties to be instantiated). In fact, this is the founding premise of causal-theoretical functionalism. Unless mental properties have causal powers, there would be little point in worrying about them. The possibility of invoking mental events in explaining behavior, or any other events, would be lost if mental properties should turn out to be causally impotent. But on the functionalist

account of mental properties, just where does a mental property get its causal powers? In particular, what is the relationship between mental property M's causal powers and the causal powers of its realizers, the Qs?

It is difficult to imagine that M's causal powers could magically materialize on their own; it is much more plausible to think—it probably is the only plausible thing to think—that M's causal powers arise out of those of its realizers, the Qs. In fact, not only do they "arise out" of them, but the causal powers of any given instance of M must be the same as those of the particular Q*i* that realizes M on that occasion. Carburetors can have no causal powers beyond those of the physical device that performs the specified function of carburetors, and an individual carburetor's causal powers must be exactly those of the particular physical device in which it is realized (if for no other reason than the simple fact that this physical device *is* the carburetor).[11] To believe that it could have excess causal powers beyond those of the physical realizer is to believe in magic: Where *could* they possibly come from? And to believe that the carburetor has fewer causal powers than the particular physical device realizing it seems totally unmotivated; just ask, "Which causal powers should we subtract?"

Let us consider this issue in some detail. On functionalism, for a psychological subject to be in mental state M is for it to be in a physical state P where P realizes M—that is, P is a physical state that is causally connected in appropriate ways with other internal physical states (some of which realize other mental states) and physical inputs and outputs. In this situation, all that there is, when the system is in mental state M, is its physical state P; being in M has no excess reality over and beyond being in P, and whatever causal powers that accrue to the system in virtue of being in M must be those of state P. It seems evident that this instance of M can have no causal powers over and beyond those of P. If my pain, here and now, is realized in a particular event of my C-fibers being stimulated, the pain must have exactly the causal powers of the particular instance of C-fiber stimulation.

But we must remember that M is multiply realized—say, by P_1, P_2, and P_3 (the finitude assumption will make no difference). If multiplicity has any meaning here, these Ps must be importantly different, and the differences that matter must be *causal* differences. To put it another way, the physical realizers of M count as different because they have different, even extremely diverse, causal powers. For this reason, it is not possible to associate a unique set of causal powers with M; each *instance* of M, of course, is an instance of P_1 or an instance of P_2 or an instance of P_3 and as such represents a specific set of causal powers, those associated with P_1, P_2, or P_3. However, M taken as a kind or property does not. That is to say, two arbitrary M-instances cannot be counted on to

have much in common in their causal powers beyond the functional causal role that defines M. In view of this, it is difficult to regard M as a property with any causal-nomological unity, and we are led to think that M has little chance of entering into significant lawful relationships with other properties. All this makes the scientific usefulness of M highly problematic.

Moreover, it has been suggested that kinds in science are individuated on the basis of causal powers; that is, to be recognized as a useful property in a scientific theory, a property must possess (or be) a determinate set of causal powers.[12] In other words, the resemblance that defines kinds in science is primarily *causal-nomological resemblance*: Things that are similar in causal powers and play similar roles in laws are classified as falling under the same kind. Such a principle of individuation for scientific kinds would seem to disqualify M and other multiply realizable properties as scientific kinds. This surely makes the science of the Ms, namely, the psychological and cognitive sciences, a dubious prospect.

These are somewhat surprising conclusions, not the least because most functionalists are ardent champions of psychology and cognitive science—in fact, of all the special sciences—as forming irreducible and autonomous domains in relation to the underlying physical-biological sciences, and this is the most influential and widely received view concerning the nature and status of psychology. We should remember that functionalism itself was largely motivated by the recognition of the multiple realizability of mental properties and a desire to protect the autonomy of psychology as a special science. The ironic thing is that if our reasoning here is not entirely off target, the conjunction of functionalism and the multiple realizability of the mental leads to the conclusion that psychology is in danger of losing its unity and integrity as a science. On functionalism, then, mental kinds are in danger of fragmenting into their multiply diverse physical realizers and ending up without the kind of causal-nomological unity and integrity expected of scientific kinds.[13]

Roles Versus Realizers: The Status of Cognitive Science

Some will object to the considerations that have led to these deflationary conclusions about the scientific status of psychological and cognitive properties and kinds as functionally conceived. Most functionalists, including many practicing cognitive and behavioral scientists, will find them surprising and unwelcome. For they believe, or want to believe, all of the following four theses: (1)

psychological-cognitive properties are multiply realizable; hence, (2) they are irreducible to physical properties; however, (3) this does not affect their status as legitimate scientific kinds; from all this it follows that (4) the cognitive and behavioral sciences form an autonomous science irreducible to more basic, "lower-level" sciences like biology and physics.

The defenders of this sort of autonomy thesis for cognitive-behavioral science will argue that the alleged fragmentation of psychological-cognitive properties as scientific properties, presented in the preceding section, was made plausible by our single-minded focus on their lower-level realizers. It is this narrow focus on the diversity of the possible realizers of mental properties that makes us lose sight of their unity as properties—the kind of unity that is invisible "bottom up." Instead, our focus should be on the "roles" that define these properties, and we should never forget that psychological-cognitive properties are "role" properties. So we might want to distinguish between "role functionalism" and "realizer functionalism."[14] Role functionalism identifies each mental property with being in a state that plays a specified causal role and keeps them clearly distinct from the physical mechanisms that fill the role, that is, the mechanisms that enable systems with the mental property to do what they are supposed to do. In contrast, realizer functionalism associates mental properties more closely with their realizers and identifies each specific *instance* of a mental property with an *instance* of its physical realizer. So the different outlooks of the two functionalisms may be stated like this:

> *Realizer Functionalism*. My experiencing pain at time *t* is identical with my C-fibers being activated at *t* (where C-fiber activation is the pain realizer in me); the octopus's experiencing pain at *t* is identical with its X-fibers being activated at *t* (where X-fiber activation is the octopus's pain realizer); and so on. The property instantiated when I experience pain at *t* is not identical with the property instantiated by the octopus when it experiences pain at *t*.

> *Role Functionalism*. My experiencing pain at time *t* is identical with my being at *t* in a state that plays causal role R (that is, the role of detecting bodily damage and triggering appropriate behavioral responses); the octopus's experiencing pain at *t* is identical with its being, at *t*, in a state that plays the same causal role R; and so on. My pain at *t* and the octopus pain at *t* share the same functional property, namely being in a state with causal role R.

Where the realizer functionalist sees differences and disunity among instances of pain, the role functionalist sees similarity and unity represented by pain's functional role. The role property associated with being in pain is what all pains have in common, and the role functionalist claims that these role properties are thought to constitute the subject matter of psychology and cognitive science; the aim of these sciences is to discover laws and regularities holding for these properties, and this can be done without attending to the physical and compositional details of their realizing mechanisms. In this sense, these sciences operate with entities and properties that are abstracted from the details of the lower-level sciences. Going back to the four theses, (1) through (4), it will be claimed that they should be understood as concerning mental properties as conceived in accordance with role functionalism.

Evidently, for role properties to serve these purposes, they must be robustly causal and nomological properties. Here is what Don Ross and David Spurrett, advocates of role functionalism, say:

> The foundational assumptions of cognitive science, along with those of other special sciences, deeply depend on role functionalism. Such functionalism is crucially supposed to deliver a kind of causal understanding. Indeed, the very point of functionalism (on role *or* realizer versions) is to capture what is salient about what systems actually do, and how they interact, *without* having to get bogged down in micro-scale physical details.[15]

These remarks on behalf of role functionalism challenge the considerations reviewed in the preceding section pointing to the conclusion that the conjunction of functionalism (in fact, role functionalism) and the multiple realizability of mental states would undermine the scientific usefulness of mental properties. The reader is urged to think about whether the remarks by Ross and Spurrett constitute an adequate rebuttal to our earlier considerations. One point the reader should notice is this: It is questionable whether, as Ross and Spurrett claim, our considerations in favor of realizer functionalism imply that we will get "bogged down in micro-scale physical details." Realizers are not necessarily, and not usually, individuated at the microphysical level.

Perhaps it might be argued that the actual practices and accomplishments of cognitive science and other special sciences go to show the emptiness of the essentially philosophical and a priori arguments of the preceding section. In spite of the heterogeneity of their underlying implementing mechanisms, functional role properties enter into laws and regularities that hold across di-

verse physical realizers. Ned Block, for example, has given some examples of psychological laws—in particular, those regarding stimulus generalization (due to the psychologist Roger Shepard)—that evidently seem to hold for all sorts of organisms and systems.[16] How these empirical results are to be correctly interpreted and understood, however, is an open question. The reader should keep in mind that an illusion of a systematic psychology and cognitive science may have been created by the fact that much of the research in these sciences focus on humans and related species. It is difficult to imagine a global scientific theory of, say, perception or memory as such, for all actual and nomologically possible psychological-cognitive systems, regardless of their modes of physical realization. A more detailed discussion of these issues takes us beyond core philosophy of mind and into the philosophy of psychology and cognitive science in a serious way. This is a good topic to reflect on for readers with an interest and background in these sciences.

For Further Reading

For statements of causal-theoretical functionalism, see David Lewis, "Psychophysical and Theoretical Identifications," and David Armstrong, "The Nature of Mind." Recommended also are Sydney Shoemaker, "Some Varieties of Functionalism," and Ned Block, "What Is Functionalism?"

Hilary Putnam, who was the first to articulate functionalism, has become one of its most severe critics; see his *Representation and Reality*, especially chapters 5 and 6. For other criticisms of functionalism, see Ned Block, "Troubles with Functionalism"; Christopher S. Hill, *Sensations: A Defense of Type Materialism*, chapter 3; and John R. Searle, *The Rediscovery of the Mind*. On the problem of qualia, see chapters 9 and 10 in this book and the suggested readings therein.

On the causal powers of functional properties, see Ned Block, "Can the Mind Change the World?"; Jaegwon Kim, *Mind in a Physical World*, chapter 2; Brian McLaughlin, "Is Role Functionalism Committed to Epiphenomenalism?"

The most influential statement of the multiple realization argument is Jerry Fodor, "Special Sciences, or the Disunity of Science as a Working Hypothesis." For the implications of multiple realization for cognitive-behavioral science, see Jaegwon Kim, "Multiple Realization and the Metaphysics of Reduction." For replies, see Ned Block, "Anti-Reductionism Slaps Back," and Jerry Fodor, "Special Sciences: Still Autonomous After All These Years." For a defense of cognitive science, see Don Ross and David Spurrett, "What to Say

to a Skeptical Metaphysician: A Defense Manual for Cognitive and Behavioral Scientists." For further discussion, see Gene Witmer, "Multiple Realizability and Psychological Laws: Evaluating Kim's Challenge."

Notes

1. This corresponds to machine functionalism's reference to the entire machine table of a Turing machine in characterizing its "internal" states. More below.

2. See David Lewis, "How to Define Theoretical Terms," and "Psychophysical and Theoretical Identifications."

3. Ramsey's original construction was in a more general setting of "theoretical" and "observational" terms rather than "psychological" and "physical-behavioral" terms. For details, see Lewis, "Psychophysical and Theoretical Identifications."

4. Here we follow Ned Block's method (rather than Lewis's) in his "What Is Functionalism?"

5. These remarks are generally in line with the "theory theory" of commonsense psychology. There is a competing account, the "simulation theory," according to which our use of commonsense psychology is not a matter of possessing a theory and applying its laws and generalizations but of "simulating" the psychology of others, using ourselves as models. See Robert M. Gordon, "Folk Psychology as Simulation," and Alvin I. Goldman, *Simulating Minds*. Prima facie, the simulation approach to folk psychology creates difficulties for the Ramsey-Lewis functionalization of mental terms. However, the precise implications of the theory need to be explored further.

6. The extension of a predicate, or concept, is the set of all things to which the predicate, or the concept, applies. So the extension of "human" is the set of all human beings. The extension of "unicorn" is the empty (or null) set.

7. For such a view, see Paul Churchland, "Eliminative Materialism and the Propositional Attitudes."

8. But it is difficult to imagine how the belief-desire-action principle *could* be shown to be empirically false. It has been argued that this principle is a priori true and hence resists empirical falsification. However, not all principles of vernacular psychology need to have the same status. It is possible that there is a core set of principles of vernacular psychology that can be considered a priori true and that suffices as a basis of the application of the Ramsey-Lewis method.

9. See Ned Block, "Troubles with Functionalism."

10. On the possibility of qualia inversion, see Sydney Shoemaker, "Inverted Spectrum"; Ned Block, "Are Absent Qualia Impossible?"; C. L. Hardin, *Color for Philosophers*; and Martine Nida-Rümelin, "Pseudo-Normal Vision: An Actual Case of Qualia Inversion?"

11. Being a carburetor is a functional property defined by a job description ("mixer of air and gasoline vapors" or some such), and a variety of physical devices can serve this purpose.

12. See, for example, Jerry Fodor, *Psychosemantics*, chapter 2.

13. For further discussion, see Jaegwon Kim, "Multiple Realization and the Metaphysics of Reduction." For replies, see Ned Block, "Anti-Reductionism Slaps Back," and Jerry Fodor, "Special Sciences: Still Autonomous After All These Years."

14. These terms are borrowed from Don Ross and David Spurrett, "What to Say to a Skeptical Metaphysician: A Defense Manual for Cognitive and Behavioral Scientists." The discussion here is indebted to this article. The distinction between role and realizer functionalism closely parallels (is identical with?) Ned Block's distinction between the functional-state identity theory and the functional specification theory in his "What Is Functionalism?" Brian McLaughlin calls realizer functionalism "filler functionalism."

15. Don Ross and David Spurrett, "What to Say to a Skeptical Metaphysician."

16. Ned Block, "Anti-Reductionism Slaps Back."

CHAPTER 7

Mental Causation

A memorable illustration of mental causation occurs in a celebrated episode in the beginning pages of Proust's *Remembrance of Things Past*. One cold, dreary winter day, the narrator's mother offers him tea, and he takes it with one of the little cakes, "petites madeleines," soaked in it. Here is what happens:

> No sooner had the warm liquid mixed with the crumbs touched my palate than a shudder ran through me and I stopped, intent upon the extraordinary thing that was happening to me. An exquisite pleasure had invaded my senses, something isolated, detached, with no suggestion of its origin. And at once the vicissitudes of life had become indifferent to me, its disasters innocuous, its brevity illusory—this new sensation having had on me the effect which love has of filling me with a precious essence.

The narrator is puzzled: Where does this sudden sense of bliss and contentment come from? Soon, a torrential rush of memories from the distant past is unleashed:

> And suddenly the memory revealed itself. The taste was that of the little piece of madeleine which on Sunday mornings at Combray (because on those mornings I did not go out before mass), when I went to say good morning to her in her bedroom, my aunt Léonie used to give me, dipping it first in her own cup of tea or tisane. . . .
>
> And as soon as I had recognized the taste of the piece of madeleine soaked in her decoction of lime-blossom which my aunt used to give me . . . immediately the old grey house upon the street, where her room

> was, rose up like a stage set to attach itself to the little pavilion opening on to the garden which had been built out behind it for my parents; and with the house the town, from morning to night and in all weathers, the Square where I used to be sent before lunch, the streets along which I used to run errands, the country roads we took when it was fine. And as in the game wherein the Japanese amuse themselves by filling a porcelain bowl with water and steeping in it little pieces of paper which until then are without character or form, but, the moment they become wet, stretch and twist and take on colour and distinctive shape, become flowers or houses or people, solid and recognizable, so in that moment all the flowers in our garden and in M. Swann's park, and the water-lilies on the Vivonne and the good folk of the village and their little dwellings and the parish church and the whole of Combray and its surroundings, taking shape and solidity, sprang into being, town and gardens alike, from my cup of tea.[1]

So begins Proust's journey into the past, in "search of lost time," which takes him more than a dozen years to complete, spanning three thousand pages. All this triggered by some madeleine crumbs soaked in a cup of tea.

This is a case of the so-called involuntary memory—where sensory or perceptual cues we encounter cause recollections of past experiences without conscious effort. It is amazing how a whiff of smell, or a tune, can bring back, totally unexpectedly, a rich panorama of images from a distant past that was apparently lost to us forever.

Returning to our philosophical concerns from the enchanting world of Proust's masterpiece, we can see in this episode several cases of causation involving mental events. The most prominent instance, one of wide literary fame, occurs when the taste of tea-soaked madeleine crumbs causes a sudden burst of recollections of the past. This is a case of mental-to-mental causation—a mental event causing another. There is also the madeleine causing our narrator to experience its taste, a case of physical-to-mental causation. The narrator first declines the tea offered by his mother, then changes his mind and takes the tea. This involves mental-to-physical causation.

When we look around, we see mental causation everywhere. In perception, objects and events around us—the computer display I am staring at, the jet passing overhead, the ocean breeze in the morning—cause visual, auditory, tactile, and other sorts of experiences. In voluntary action, our desires and intentions cause our limbs to move so as to rearrange the objects around us. On a grander scale, it is human knowledge, wishes, dreams, greed, passions, and inspirations that led our forebears to build the pyramids of Egypt and the

Great Wall of China, and to create the glorious music, literature, and artworks that form our cultural heritage. These mental capacities and functions have also been responsible for nuclear weapons, global warming, disastrous oil spills, and the destruction of the rain forests. Mental phenomena are intricately and seamlessly woven into the complex mosaic of causal relations of the world. Or so it seems, at least.

If your mind is going to cause your limbs to move, it presumably must first cause an appropriate neural event in your brain. But how is that possible? How can a mind, or a mental phenomenon, cause a bundle of neurons to fire? Through what mechanisms does a mental event, like a thought or a feeling, manage to initiate, or insert itself into, a causal chain of electrochemical neural events? And how is it possible for a chain of physical and biological events and processes to burst, suddenly and magically, into a full-blown conscious experience, with all its vivid colors, shapes, smells, and sounds? Think of your total sensory experience right now—visual, tactual, auditory, olfactory, and the rest: How is it possible for all this to arise out of molecular activities in the gray matter of your brain?

Agency and Mental Causation

An agent is someone with the capacity to *perform actions for reasons*, and most actions involve bodily movements. We are all agents in that sense: We do such things as turning on the stove, heating water in a kettle, making coffee, and entertaining friends. An action is something we "do"; it is unlike a "mere happening," like sweating, running a fever, or being awakened by the noise of a jackhammer. These are what happens to us; they are not in our control. Implicit in the notion of action is the idea that an agent is in control of what she does, and the control here can only mean causal control.

Let us look into this in some detail. Consider Susan's heating water in a kettle. This must at least include her *causing* the water in the kettle to rise in temperature. Why did Susan heat the water? When someone performs an action, it always makes sense to ask why, even if the correct answer may be "For no particular reason." Susan, let us suppose, heated water to make tea. That is, she *wanted* to make tea and *believed* that she needed hot water to do that—and to be boringly detailed, she *believed* that by heating water in the kettle she could get the hot water she needed. When we know all this, we know why Susan heated water; we understand her action. Beliefs and desires guide actions, and by citing appropriate beliefs and desires, we are able to explain and make sense of why people do what they do.[2]

We may consider the following statement as the fundamental principle that connects desire, belief, and action:

> *Desire-Belief-Action Principle.* (DBA): If agent S desires something and believes that doing A is an optimal way of securing it, S will do A.

As stated, DBA is too strong. For one thing, we often choose not to act on our desires, and sometimes we change them, or try to get rid of them, when we realize that pursuing them is too costly and may lead to consequences that we want to avoid. For example, you wake up in the middle of the night and want a glass of milk, but the thought of getting out of bed in the chilly winter night and going down two long flights of stairs to the dark kitchen talks you out of it. Further, even when we are ready to act on our desires and beliefs, we may find ourselves physically unable to perform the action: It may be that when you have finally overcome your aversion to getting out of the bed, you find yourself chained to the bedposts!

To save DBA, we can tinker with it in various ways; for example, we can add further conditions to the antecedent of DBA (such as that there are no other conflicting desires) or weaken the consequent (for example, by turning it into a probability or tendency statement, or adding the all-purpose hedge "other things being equal" or "under normal conditions"). In any event, there seems little question that a principle like DBA is fundamental to the way we explain and understand actions, both our own and those of others around us. DBA is often taken to be the fundamental schema that anchors reason-based explanations of actions, or "rationalizations." In saying this, we need not imply that beliefs and desires are the only possible reasons for actions; for example, emotions and feelings are often invoked as reasons, as witness, "I hit him because he insulted my wife and that made me angry," or, "He jumped up and down for joy."[3]

What the exceptions to DBA we have considered show is that an agent may have a reason—a "good" reason—to do something but fail to do it. Sometimes there may be more than one belief-desire pair that is related to a given action, as specified by DBA: In addition to your desire for a glass of milk, you heard a suspicious noise from downstairs and wanted to check it out. Let us suppose that you finally did get out of bed to venture down the stairway. Why did you do that? What explains it? It is possible that you went downstairs because you thought you really ought to check out the noise, not out of your desire for milk. If so, it is your desire to check the noise, not your desire for milk, that explains why you went downstairs in the middle of the night. It would be

correct for you to say, "I went downstairs because I wanted to check out the noise," but incorrect to say, "I went downstairs because I wanted a glass of milk," although you did get your milk too. We can also put the point this way: Your desire to check out the noise and your desire for milk were both *reasons*—in fact, *good reasons*—for going downstairs, but the first, not the second, was the *reason for which* you did what you did; it was the *motivating reason*. And it is "reason for which," not mere "reason for," that explains the action. But what precisely is the difference between them? That is, what distinguishes explanatory reasons from reasons that do no explanatory work?

A widely accepted—though by no means undisputed—answer defended by Donald Davidson is the simple thesis that a reason for which an action is done is one that *causes* it.[4] That is, what makes a reason for an action an explanatory reason is its role in the causation of that action. Thus, on Davidson's view, the crucial difference between my desire to check out the noise and my desire for a glass of milk lies in the fact that the former, not the latter, caused me to go downstairs. This makes explanation of action by reasons, or "rationalizing" explanation, a species of causal explanation: Reasons explain actions in virtue of being their causes.

If this is correct, it follows that agency is possible only if mental causation is possible. For an agent is someone who is able to act for reasons and whose actions can be explained and evaluated in terms of the reasons for which she acted. This entails that reasons—that is, mental states like beliefs, desires, and emotions—must be able to cause us to do what we do. Since what we do almost always involves movements of our limbs and other bodily parts, this means that agency—at least human agency—presupposes the possibility of mental-to-physical causation. Somehow your beliefs and desires cause your limbs to move in appropriate ways so that in ten seconds you find your whole body, made up of untold billions of molecules and weighing over a hundred pounds, displaced from your bedroom to the kitchen. A world in which mental causation does not exist is one in which there are no agents and no actions.

Mental Causation, Mental Realism, and Epiphenomenalism

Perception involves the causation of mental events—perceptual experiences and beliefs—by physical processes. In fact, the very idea of perceiving something—say, seeing a tree—involves the idea that the object seen is a cause of your visual experience. Suppose that there is a tree in front of you and that you are having a visual experience of the sort you would be having if your retinas were stimulated

by the light rays reflected by the tree. But you would not be seeing the tree if a holographic image of a tree, visually indistinguishable from the tree, were interposed between you and the tree. You would be seeing the holographic image of a tree, not the tree, even though your perceptual experience in the two cases would have been exactly alike. Evidently, this difference too is a causal one: Your visual experience is caused by a tree holograph, not by the tree.

Perception is our sole window on the world; without it, we could learn nothing about what goes on around us. If, therefore, perception necessarily involves mental causation, there could be no knowledge of the world without mental causation. Moreover, a significant part of our knowledge of the world is based on experimentation, not mere observation. Experimentation differs from passive observation in that it requires our active intervention in the course of natural events; we design and deliberately set up the experimental conditions and then observe the outcome. This means that experimentation presupposes mental-to-physical causation and is impossible without it. Much of our knowledge of causal relations—in general, knowledge of what happens under what conditions—is based on experimentation, and such knowledge is essential not only to our theoretical understanding of the world but also to our ability to predict and control the course of natural events. We must conclude, then, that if minds were not able to causally connect with physical events and processes, we could have neither the practical knowledge required to inform our decisions and actions nor the theoretical knowledge that gives us an understanding of the world around us.

Mental-to-mental causation also seems essential to human knowledge. Consider the process of inferring one proposition from another. Suppose someone asks you, "Is the number of planets odd or even?" If you are like most people, you would probably proceed like this: "Well, how many planets are there? Eight, of course, and eight is an even number because it is a multiple of two. So the answer is: The number is even." You have just inferred the proposition that there are an even number of planets from the proposition that there are eight planets, and you have formed a new belief based on this inference. This process evidently involves mental causation: Your belief that the number of planets is even was caused, through a chain of inference, by your belief that there are eight planets. Inference is one way in which beliefs generate other beliefs. A brief reflection makes it evident that most of our beliefs are generated by other beliefs we hold, and "generation" here could only mean causal generation. It follows, then, that all three types of mental causation—mental-to-physical, physical-to-mental, and mental-to-mental—are implicated in the possibility of human knowledge.

Epiphenomenalism is the view that although all mental events are caused by physical events, they are only "epiphenomena"—that is, events without powers to cause any other event. Mental events are effects of physical (presumably neural) processes, but they do not in turn cause anything else, being powerless to affect physical events or even other mental events; they are the absolute termini of causal chains. The noted nineteenth-century biologist T. H. Huxley has this to say about the consciousness of animals:

> The consciousness of brutes would appear to be related to the mechanism of their body simply as a collateral product of its working and to be as completely without any power of modifying that working as the steam-whistle which accompanies the work of a locomotive engine is without influence upon its machinery. Their volition, if they have any, is an emotion indicative of physical changes, not a cause of such changes.

What about human consciousness? Huxley goes on:

> It is quite true that, to the best of my judgment, the argumentation which applies to brutes holds equally good of men; and, therefore, that all states of consciousness in us, as in them, are immediately caused by molecular changes of the brain-substance. It seems to me that in men, as in brutes, there is no proof that any state of consciousness is the cause of change in the motion of the matter of organism. . . . We are conscious automata.[5]

What was Huxley's argument that convinced him that the consciousness of animals is causally inert? Huxley's reasoning appears to have been something like this: In animal experiments (Huxley mentions experiments with frogs), it can be shown that animals are able to perform complex bodily operations when we have compelling neuroanatomical evidence that they cannot be conscious, and this shows that consciousness is not needed as a cause of these bodily behaviors. Moreover, similar phenomena are observed in cases involving humans: As an example, Huxley cites the case of a brain-injured French sergeant who was reduced to a condition comparable to that of a frog with the anterior part of its brain removed—that is, we have ample anatomical reason to believe that the unfortunate war veteran had no capacity for consciousness—but who could perform complex actions of the kind that we normally think require consciousness, like avoiding obstacles when walking around in a familiar place, eating and drinking, dressing and undressing, and going to bed at the accustomed time. Huxley takes cases of this kind as a basis for his claim that consciousness is not a

cause of behavior production in animals or humans. Whether Huxley's reasoning is sound is something to think about.[6]

Consider a moving car and the series of shadows it casts as it races along the highway: The shadows are caused by the moving car but have no effect on the car's motion. Nor are the shadows at different times causally connected: The shadow at a given instant *t* is caused not by the shadow an instant earlier but by the car itself at *t*. A person who observes the moving shadows but not the car may be led to attribute causal relations between the shadows, the earlier ones causing the later ones, but he would be mistaken. Similarly, you may think that your headache has caused your desire to take aspirin, but that, according to the epiphenomenalist, would be a similar mistake: The headache and the desire for aspirin are both caused by two successive states of the brain, but they are not related as cause to effect any more than two successive shadows of the moving car. The apparent regularities that we observe in mental events, the epiphenomenalist argues, do not represent genuine causal connections; like the regularities characterizing the car's moving shadows or the successive symptoms of a disease, they are merely reflections of the real causal processes at a more fundamental level.

These are the claims of epiphenomenalism. Few philosophers have been self-professed epiphenomenalists, although there are those whose views appear to lead to such a position (as we will see below). We are more likely to find epiphenomenalist thinking among scientists in brain science. At least, some scientists seem to treat mentality, especially consciousness, as a mere shadow or afterglow thrown off by the complex neural processes going on in the brain; these physical-biological processes are what at bottom do all the pushing and pulling to keep the human organism functioning. If conscious events really had causal powers to influence neural events, there could be no complete neural-physical explanations of neural events unless consciousness was explicitly brought into neuroscience as an independent causal agent in its own right. That is, there could be no complete physical-biological theory of neural phenomena. It would seem that few neuroscientists would countenance such a possibility. (For further discussion, see chapter 10.)

How should we respond to the epiphenomenalist stance on the status of mind? Samuel Alexander, a leading emergentist during the early twentieth century, comments on epiphenomenalism with a pithy remark:

> [Epiphenomenalism] supposes something to exist in nature which has nothing to do, no purpose to serve, a species of noblesse which depends on the work of its inferiors, but is kept for show and might as well, and undoubtedly would in time, be abolished.[7]

Alexander is saying that if epiphenomenalism is true, mind has no work to do and hence is entirely useless, and it is pointless to recognize it as something real. Our beliefs and desires would have no role in causing our decisions and actions and would be entirely useless in their explanations; our perception and knowledge would have nothing to do with our artistic creations or technological inventions. *Being real and having causal powers go hand in hand; to deprive the mind of causal potency is in effect to deprive it of its reality.*

It is important to see that this is not an *argument* against epiphenomenalism: Alexander only points out, in a stark and forceful way, what accepting epiphenomenalism would entail. We should also remind ourselves that the typical epiphenomenalist does not reject the reality of mental causation altogether; she only denies mind-to-body and mind-to-mind causation, not body-to-mind causation. In this sense, she gives the mental a well-defined place in the causal structure of the world; mental events are integrated into that structure as effects of neural processes. This suggests that there is a stronger form of epiphenomenalism, according to which the mental is both causeless and effectless—that is, the mental is simply noncausal. To a person holding such a view, mental events are in total causal isolation from the rest of the world, even from other mental events; each mental event is a solitary island, with no connection to anything else. (Recall the discussion of the causal status of immaterial substances, in chapter 2.) Its existence would be entirely inexplicable since it has no cause, and it would make no difference to anything else since it has no effect. It would be a mystery how the existence of such things could be known to us. As Alexander declares, they could just as well be "abolished"—that is, regarded as nonexistent. No philosopher appears to have explicitly held or argued for this stronger form of epiphenomenalism; however, as we will see, there are views on the mind-body problem that seem to lead to a radical epiphenomenalism of this kind.

So why not grant the mind full causal powers, among them the power to influence bodily processes? This would give the mental a full measure of reality and recognize what after all is so manifestly evident to common sense. That is just what Descartes tried to do with his thesis that minds and bodies, even though they are substances of very different sorts, are in intimate causal commerce with each other. But we have seen what seem like impossible difficulties besetting his program (chapter 2).

Everyone will acknowledge that mental causation is a desideratum—something important to save. Jerry Fodor is not jesting when he writes:

> I'm not really convinced that it matters very much whether the mental is physical; still less that it matters very much whether we can prove that it is.

> Whereas, if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying . . . , if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world.[8]

For Fodor, then, mental causation is absolutely nonnegotiable. And it is understandable why anyone should feel this way: Giving up mental causation amounts to giving up our conception of ourselves as agents and cognizers. Is it even *possible* for us to give up the idea that we are agents who decide and act, that we perceive and know certain things about the world? Can we live our lives as epiphenomenalists? That is, as "practicing" epiphenomenalists?

In his first sentence in the foregoing quoted passage, Fodor is saying that being able to defend a theory of the mind-body relation is far less important than safeguarding mental causation. That is not an implausible perspective to take: Whether a stance on the mind-body problem is acceptable depends importantly, if not solely, on how successful it is in giving an account of mental causation. On this criterion, Descartes's substance dualism, in the opinion of many, must be deemed a failure. So the main question is this: Which positions on the mind-body problem allow full-fledged mental causation and provide an explanation of how it is possible? We consider this question in the sections to follow.

Psychophysical Laws and "Anomalous Monism"

The expulsion of Cartesian immaterial minds perhaps brightens the prospect of understanding how mental causation is possible. For we no longer have to contend with a seemingly hopeless question: How could immaterial souls with no physical characteristics—no bulk, no mass, no energy, no charge, and no location in space—causally influence, and be influenced by, physical objects and processes? Today few, though not all, philosophers or scientists regard minds as substances of a special nonphysical sort; mental events and processes are now viewed as occurring in complex physical systems like biological organisms, not in immaterial minds. The problem of mental causation, therefore, is now formulated in terms of two kinds of events, mental and physical, not in terms of two kinds of substances: How is it possible for a mental event (such as a pain or a thought) to cause a physical event (a limb withdrawal, an utterance)? Or in terms of properties: How is it possible for an

instantiation of a mental property (for example, that of experiencing pain) to cause a physical property to be instantiated?

But why is this supposed to be a "problem"? We do not usually think that there is a special philosophical problem about, say, chemical events causally influencing biological processes or a nation's economic and political conditions causally affecting each other. So what is it about mentality and physicality that make causal relations between them a philosophical problem? For substance dualism, it is, at bottom, the extreme heterogeneity of minds and bodies, in particular, the nonspatiality of minds and the spatiality of bodies (as argued in chapter 2), that makes causal relations between them problematic. Given that mental substances have now been expunged, aren't we home free with mental causation? The answer is that certain other assumptions and doctrines that demand our respect present apparent obstacles to mental causation.

One such doctrine centers on the question of whether there are *laws connecting mental phenomena with physical phenomena*—that is, psychophysical laws—that are thought to be needed to underwrite causal connections between them. Donald Davidson's well-known "anomalism of the mental" states that there can be no such laws.[9] A principle connecting laws and causation that is widely, though not universally, accepted, is this: *Causally connected events must instantiate, or be subsumed under, a law*. If heating a metallic rod causes its length to increase, there must be a law connecting events of the first type and events of the second type; that is, there must be a law stating that the heating of a metallic rod is followed by an increase in its length. But if causal connections require laws and there are no laws connecting mental events with physical events, it would seem to follow that there could be no mental-physical causation. This line of reasoning is examined in more detail later in the chapter. But is there any reason to doubt the existence of laws connecting mental and physical phenomena?

In earlier chapters, we often assumed that there are lawful connections between mental and physical events; you surely recall the stock example of pain and C-fiber excitation. The psychoneural identity theory, as we saw, assumes that each type of mental event is lawfully correlated with a type of physical event. Talk of "physical realization" of mental events also presupposes that there are lawlike connections between a mental event of a given kind and its diverse physical realizers, for a physical realizer of a mental event must at least be sufficient, as a matter of law, for the occurrence of that mental event. The very idea of a "neural correlate" seems to imply that there are psychophysical laws; if a mental state and its neural correlate co-occur, that has to be a lawlike relationship, not an accidental connection. Davidson explicitly restricts his

claim about the nonexistence of psychophysical laws to intentional mental events and states ("propositional attitudes")—that is, states with propositional content, like beliefs, desires, hopes, and intentions; he is not concerned with sensory events and states, like pains, visual sensings of color, and mental images. Why does Davidson think that no laws exist connecting, say, beliefs with physical-neural events? Doesn't every mental event have a neural substrate, that is, a neural state that, as a matter of law, suffices for its occurrence?

Before we take a look at Davidson's argument, let us consider some examples. Take the belief that it is unseemly for the president of the United States to get a $500 haircut. How reasonable is it to expect to find a neural substrate for this belief? Is it at all plausible to think that all and only people who have this belief share some specific neural state? It makes perfectly good sense to try to find neural correlates for pains, sensations of thirst and hunger, visual images, and the like, but somehow it does not seem to make much sense to look for the neural correlates of mental states like our sample belief, or for things like your sudden realization that you have a philosophy paper due in two days, your hope that airfares to California will come down after Christmas, and the like. Is it just that these mental states are so complex that it is very difficult, perhaps impossible, for us to discover their neural bases? Or is it the case that they are simply not the sort of state for which neural correlates could exist and that it makes no sense to look for them?

This is not intended as an argument for the impossibility of psychophysical laws but it should dispel, or at least weaken, the strong presumption many of us are apt to hold that there must "obviously" be psychophysical laws since mentality depends on what goes on in the brain. It is now time to turn to Davidson's famous but notoriously difficult argument against psychophysical laws.[10]

A crucial premise of Davidson's argument is the thesis that the ascription of intentional states, like beliefs and desires, is regulated by certain *principles of rationality* that ensure that the total set of such states attributed to a person will be as rational and coherent as possible. This is why, for example, we refrain from attributing to a person manifestly contradictory beliefs, even when the sentences uttered have the surface logical form of a contradiction. When someone replies, "Well, I do and I don't," when asked, "Do you like Ralph Nader?" we do not take her to be expressing a literally contradictory belief—the belief that she both likes and does not like Nader. Rather, we take her to be saying something like, "I like some aspects of Nader (say, his concerns for social and economic justice), but I don't like other aspects (say, his presidential ambitions)." If she were to insist, "No, I don't mean that; I really both do and don't like Nader, period," we would not know what to make of her; perhaps

her "and" does not mean what the English "and" means, or perhaps she does not have a full grasp of "not." We cast about for some consistent interpretation of her meaning because an interpreter of a person's speech and mental states is under the mandate that an acceptable interpretation must make her come out with a reasonably coherent set of beliefs—as coherent and rational as evidence permits. When no minimally coherent interpretation is possible, we are apt to blame our own interpretive efforts rather than accuse our subject of harboring explicitly inconsistent beliefs. We also attribute to a subject beliefs that are obvious logical consequences of beliefs already attributed to him. For example, if we have ascribed to a person the belief that Boston is less than sixty miles from Providence, we would, and should, ascribe to him the belief that Boston is less than seventy miles from Providence, the belief that Boston is less than one hundred miles from Providence, and countless others. We do not need independent evidence for these further belief attributions; if we are not prepared to attribute any one of these further beliefs, we should reconsider our original belief attribution and be prepared to withdraw it. Our concept of belief does not allow us to say that someone believes that Boston is within sixty miles of Providence but does not believe that it is within seventy miles—unless we are able to give an intelligible explanation of how this could happen in this particular case. This principle, which requires that the set of beliefs be "closed" under obvious logical entailment, goes beyond the simple requirement of consistency in a person's belief system; it requires that the belief system be coherent as a whole—it must in some sense hang together, without unexplainable gaps. In any case, Davidson's thesis is that the requirement of rationality and coherence[11] is of the essence of the mental—that is, it is *constitutive* of the mental in the sense that it is exactly what makes the mental mental. Keep in mind that Davidson is speaking only of intentional states, like belief and desire, not sensory states and events like pains and afterimages. (For further discussion, see chapter 8 on interpretation theory.)

But it is clear that the physical domain is subject to no such requirement; as Davidson says, the principle of rationality and coherence finds "no echo" in physical theory. Suppose now that we have laws connecting beliefs with brain states; in particular, suppose we have laws that specify a neural substrate for each of our beliefs—a series of laws of the form "N occurs to a person at *t* if and only if B occurs to that person at *t*," where N is a neural state and B is a belief. If such laws were available, we could attribute beliefs to a subject, *one by one*, independently of the constraints of the rationality principle. For in order to determine whether she has a certain belief B, all we would need to do is ascertain whether B's neural substrate N is present in her; there would be no

need to check whether this belief makes sense in the context of her other beliefs or even what other beliefs she has. In short, we could read her mind by reading her brain. The upshot is that the practice of belief attribution would no longer be regulated by the rationality principle. By being connected by law with neural state N, belief B becomes hostage to the constraints of physical theory. On Davidson's view, as we saw, the rationality principle is constitutive of mentality, and beliefs that have escaped its jurisdiction can no longer be considered beliefs. If, therefore, belief is to retain its identity and integrity as a mental phenomenon, its attribution must be regulated by the rationality principle and hence cannot be connected by law to a physical substrate.

Let us assume that Davidson has made a plausible case for the impossibility of psychophysical laws (we may call his thesis "psychophysical anomalism") so that it is worthwhile to explore its consequences. One question that was raised earlier is whether it might make mental causation impossible. Here the argument could go like this: Causal relations require laws, and this means that causal relations between mental events and physical events require psychophysical laws, laws connecting mental and physical events. But Davidson's psychophysical anomalism holds that there can be no such laws, whence it would appear to follow that there can be no causal relations between mental and physical phenomena. Davidson, however, is a believer in mental causation; he explicitly holds that mental events sometimes cause, and are caused by, physical events. This means that Davidson must reject the argument just sketched that attempts to derive the nonexistence of mental causation from the nonexistence of psychophysical laws. How can he do that?

What Davidson disputes in this argument is its first step, namely, the inference from the premise that *causation requires laws* to the conclusion that *psychophysical causation requires psychophysical laws*. Let us look into this in some detail. To begin, what is it for one individual event *c* to cause another individual event *e*? This holds, on Davidson's view, only if the two events instantiate a law, in the following sense: *c* falls under a certain event kind (or description) F, *e* falls under an event kind G, and there is a law connecting events of kind F with events of kind G (as cause to effect). This is a form of the influential nomological account of causation: Causal connections must instantiate, or be subsumed under, general laws. Suppose, then, that a particular mental event, *m*, causes a physical event, *p*. This means, according to the nomological conception of causation, that for some event kinds, C and E, *m* falls under C and *p* falls under E, and there is a law that connects events of kind C with events of kind E. This makes it evident that laws connect individual events only as they fall under kinds. Thus, when psychophysical anomal-

ism says that there are no psychophysical laws, what it says is that there are no laws connecting mental kinds with physical kinds. So what follows is only that *if mental event* m *causes physical event* p, *kinds C and E, under which* m *and* p, *respectively*, *fall and which are connected by law, must both be physical kinds*. That is to say, a purely physical law must underwrite this causal relation. In particular, this means that C, under which mental event *m* falls, cannot be a mental kind; it must be a physical one. From which it follows that *m* is a physical event! For an event is mental or physical according to whether it falls under a mental kind or a physical kind. Note that this "or" is not exclusive; *m*, being a mental event, must fall under a mental kind, but that does not prevent it from falling under a physical kind as well. This argument applies to all mental events that are causally related to physical events, and there appears to be no reason not to think that every mental event has some causal connection, directly or via a chain of other events, with a physical event. All such events, on Davidson's argument, are physical events.[12]

That is Davidson's "anomalous monism." It is a monism because it claims that all individual events, mental events included, are physical events (you will recall this as "token physicalism"; see chapter 1). Moreover, it is physical monism that does not require psychophysical laws; in fact, as we just saw, the argument for it requires the nonexistence of such laws, whence the term "anomalous" monism. Davidson's world, then, looks like this: It consists exclusively of physical objects and physical events, but some physical events fall under mental kinds (or have mental descriptions) and therefore are mental events. Laws connect physical kinds and properties with other physical kinds and properties, and these laws generate causal relations between individual events. Thus, all causal relations of this world are exclusively grounded in physical laws.

Is Anomalous Monism a Form of Epiphenomenalism?

One of the premises from which Davidson derives anomalous monism is the claim that mental events can be, and sometimes are, causes and effects of physical events. On anomalous monism, however, to say that a mental event *m* is a cause of an event *p* (*p* may be mental or physical) amounts only to this: *m* has a physical property Q (or falls under a physical kind Q) such that an appropriate law connects Q (or events with property Q) with some physical property P of *p*. Since no laws exist that connect mental and physical properties, purely physical laws must do all the causal work, and this means that individual events can enter into causal relations only because they possess physical properties that

figure in laws. Consider an example: Your desire for a drink of water causes you to turn on the tap. On Davidson's nomological conception of causation, this requires a law that subsumes the two events, your desiring a drink of water and your turning on the tap. However, psychophysical anomalism says that this law must be a physical law, since there are no laws connecting mental kinds with physical kinds. Hence, your desire for a drink of water must be redescribed physically—that is, a suitable physical property of your desire must be identified—before it can be brought under a law. In the absence of psychophysical laws, therefore, it is the physical properties of mental events that determine, wholly and exclusively, what causal relations they enter into. In particular, the fact that your desire for a drink of water is a desire for a drink of water—that is, the fact that it is an event of this mental kind—apparently has no bearing on its causation of your turning on the tap. What is causally relevant is its physical properties—presumably the fact that it is a neural, or physicochemical, event of a certain kind.

It seems, then, that under anomalous monism, mental properties are causal idlers with no work to do. To be sure, anomalous monism is not epiphenomenalism in the classic sense, since individual mental events are allowed to be causes of other events. The point, though, is that it is an epiphenomenalism of *mental properties*—we may call it "mental property epiphenomenalism"[13]—in that it renders mental properties and kinds causally irrelevant. Moreover, it is a form of radical epiphenomenalism described earlier: Mental properties play no role in making mental events either causes or effects. To make this vivid: If you were to redistribute mental properties over the events of this world any way you please—you might even remove them entirely from all events, making all of them purely physical—that would not alter, in the slightest way, the network of causal relations of this world; it would not add or subtract a single causal relation anywhere in the world!

This shows the importance of properties in the debate over mental causation: It is the causal efficacy of mental properties that we need to vindicate and give an account of. With mental substances out of the picture, there are only mental properties left to play any causal role, whether these are construed as properties of events or of objects. If mentality is to do any causal work, it must be the case that having a given mental property rather than another, or having it rather than not having it, must make a causal difference; it must be the case that an event, because it has a certain mental property (for example, being a desire for a drink of water), enters into a causal relation (it causes you to look for a water fountain) that it would otherwise not have entered into. We must

therefore conclude that Davidson's anomalous monism fails to pass the test of mental causation; by failing to account for the causal efficacy and relevance of mental properties, it fails to account for the possibility of mental causation.

The challenge posed by Davidson's psychophysical anomalism, therefore, is to answer the following question: How can anomalous mental properties, properties that are not fit for laws, be causally efficacious properties? It would seem that there are only two ways of responding to this challenge: First, we may try to reject its principal premise, namely, psychophysical anomalism, by finding faults with Davidson's argument and then offering plausible reasons for thinking that there are indeed psychophysical laws that can underwrite psychophysical causal relations. Second, we may try to show that the nomological conception of causality—in particular, as it is construed by Davidson—is not the only way to understand causation and that there are alternative conceptions of causation on which mental properties, though anomalous, could still be causally efficacious. Let us explore the second possibility.

Counterfactuals to the Rescue?

There indeed is an alternative approach to causation that on its face does not seem to require laws, and this is the counterfactual account of causation. On this approach, to say that event *c* caused event *e* is to say that if *c* had not occurred, *e* would not have occurred.[14] The thought that a cause is the *sine qua non* condition, or *necessary* condition, of its effect is a similar idea. This approach has much intuitive plausibility. The overturned space heater caused the house fire. What makes it so? Because if the space heater had not overturned, the fire would not have occurred. What is the basis of saying that the accident was caused by a sudden braking on a rain-slick road? Because if the driver had not suddenly stepped on his brake pedal on the wet road, the accident would not have occurred. In such cases we seem to depend on counterfactual ("what if") considerations rather than laws. Especially if you insist on exceptionless "strict" laws, as Davidson does, we obviously are not in possession of such laws to support these perfectly ordinary and familiar causal claims, claims that we regard as well supported.

The situation seems the same when mental events are involved: There is no mystery about why I think that my desire for a drink of water caused me to step into the dark kitchen last night and stumble over the sleeping dog. It's because of the evidently true counterfactual "If I had not wanted a drink of water last night, I would not have gone into the kitchen and stumbled over the

dog." In confidently making these ordinary causal or counterfactual claims, we seem entirely unconcerned about the question of whether there are laws about wanting a glass of water and stumbling over a sleeping dog. Even if we were to reflect on such questions, we would be undeterred by the unlikely possibility that such laws exist or can be found. To summarize, then, the idea is this: We know that mental events, in virtue of their mental properties, can, and sometimes do, cause physical events because we can, and sometimes do, know appropriate mental-physical counterfactuals to be true. Mental causation is possible because such counterfactuals are sometimes true.

The counterfactual account of causation opens up the possibility of explaining mental causation in terms of how mental-physical counterfactuals can be true. To show that there is a special problem about mental causation, you must show that there are problems about these counterfactuals.

So are there special problems about these psychophysical counterfactuals? Do we have an understanding of how such counterfactuals can be true? There are many philosophical puzzles and difficulties surrounding counterfactuals, especially about their "semantics"—that is, conditions under which counterfactuals can be evaluated as true or false. There are two main approaches to counterfactuals: (1) the nomic-derivational approach, and (2) the possible-world approach.[15] On the nomic-derivational approach, the counterfactual conditional "If P were the case, Q would be the case" (where P and Q are propositions) is true just in case the consequent, Q, of the conditional can be logically derived from its antecedent, P, when taken together with laws and statements of conditions holding on the occasion.[16] Consider an example: "If this match had been struck, it would have lighted." This counterfactual is true since its consequent, "The match lighted," can be derived from its antecedent, "The match was struck," in conjunction with the law "Whenever a dry match is struck in the presence of oxygen, it lights," taken together with the auxiliary premises "The match was dry" and "There was oxygen present."

It should be immediately obvious that on this analysis of counterfactuals, the counterfactual account of mental causation does not make the problem of mental causation go away. For the truth of the psychophysical counterfactuals—like "If I had not wanted to check out the strange noise, I would not have gone downstairs," and, "If Jones's C-fibers had been activated, she would have felt pain"—would require laws that would enable the derivation of the physical consequents from their psychological antecedents (or vice versa), and this evidently requires psychophysical laws, laws connecting mental with physical phenomena. On the nomic-derivational approach, therefore, Davidson's problem of psychophysical laws arises again.

Let us consider then the possible-world approach to the truth conditions of counterfactuals. In a simplified form, it says this: The counterfactual "If P were the case, Q would be the case" is true just in case Q is true in the world in which P is true and that, apart from P's being true there, is as much like the actual world as possible. (To put it another way: Q is true in the "closest" P-world.)[17] To ascertain whether this counterfactual is true, we go through the following steps: Since this is a counterfactual, its antecedent, P, is false in the actual world. We must go to a possible world ("world" for short) in which P is true and see whether Q is also true there. But there are many worlds in which P is true—that is, there are many P-worlds—and in some of these Q is true and in others false. So which P-world should we pick in which to check on Q? The answer: Pick the P-world that is the most similar, or the "closest," to the actual world. The counterfactual "If P were true, Q would be true" is true if Q is true in the closest P-world; it is false if Q is false in that world.

Let us see how this works with the counterfactual "If this match had been struck, it would have lighted." In the actual world, the match was not struck; so suppose that the match was struck (this means, go to a world in which the match was struck), but keep other conditions the same as much as possible. Certain other conditions must be altered under the counterfactual supposition that the match was struck: For example, in the actual world the match lay motionless in the matchbox and there was no disturbance in the air in its vicinity, so these conditions have to be changed to keep the world consistent as a whole. However, we need not, and should not, change the fact that the match was dry and the fact that sufficient oxygen was present in the ambient air. So in the world we have picked, the following conditions, among others, obtain: The match was struck, it was dry, and oxygen was present in the vicinity. The counterfactual is true if and only if the match lighted in that world. Did the match light in that world? In asking this question, we are asking which of the following two worlds is closer to the actual world:[18]

W_1: The match was struck; it was dry; oxygen was present; the match lighted.
W_2: The match was struck; it was dry; oxygen was present; the match did not light.

We would judge, it seems, that of the two, W_1 is closer to the actual world, thereby making the counterfactual come out true. But why do we judge this way?

There seems to be only one answer: Because in the actual world there is a lawful regularity to the effect that when a dry match is struck in the presence

of oxygen, it ignites, and this law holds in W_1, but not W_2. That is why W_1 is closer to the actual world than W_2 is. So in judging that this match, which in fact was dry and bathed in oxygen, would have lighted if it had been struck, we seem to be making crucial use of the law just mentioned. If in the actual world dry matches, when struck in the presence of oxygen, seldom or never light, there seems little question that we would go for W_2 as the closer world and judge the counterfactual "If this match had been struck, it would have lighted" to be false. If this is right, the counterfactual model of causation does not entirely free us from laws, as we had hoped; it seems that at least in some cases we must still resort to laws and lawful regularities.

Now consider a psychophysical counterfactual: "If Brian had not wanted to check out the noise, he wouldn't have gone downstairs." Suppose that we take this counterfactual to be true, and on that basis we judge that Brian's desire to check out the noise caused him to go downstairs. Consider the following two worlds:

W_3: Brian didn't want to check out the noise; he didn't go downstairs.
W_4: Brian didn't want to check out the noise; he went downstairs anyway.

If W_4 is closer to the actual world than W_3 is, that would falsify our counterfactual. So why should we think that W_3 is closer than W_4? In the actual world, Brian wanted to check out the noise and went downstairs. As far as these two particular facts are concerned, W_4 evidently is closer to the actual world than W_3 is. So why do we hold W_3 to be closer and hence the counterfactual to be true? The only plausible answer, again, seems to be something like this: We know, or believe, that there are certain lawful regularities and propensities governing Brian's wants, beliefs, and so on, on the one hand, and his behavior patterns, on the other, and that, given the absence of something like a desire to check out a suspicious noise, along with other conditions prevailing at the time, his not going downstairs at that particular time fits these regularities and propensities better than the supposition that he would have gone downstairs at that time. We consider such regularities and propensities, that is, facts about a person's personality, to be reliable and lawlike and commonly appeal to them in assessing counterfactuals of this kind (and also in making predictions and guesses as to how a person will behave), even though we may have only the vaguest idea about the details and lack the ability to articulate them in a precise way.

Again, the relevance of psychophysical laws to mental causation is apparent. Although there is room for further discussion, it is plausible that considerations of lawful regularities governing mental and physical phenomena often seem crucially involved in the evaluation of psychophysical counterfactuals of the sort that can ground causal relations. We need not know the details of such regularities, but we must believe that they exist and know their rough content and shape to be able to evaluate these counterfactuals as true or false. So are we back where we started, with Davidson and his argument for the impossibility of psychophysical laws?

Not exactly, fortunately. The laws involved in evaluating counterfactuals, as is clear from our examples, need not be laws of the kind Davidson has in mind—what he calls "strict" laws. These are exceptionless, explicitly articulated laws that form a closed and comprehensive theory, like the laws of physics. Rather, the laws involved in evaluating these quotidian counterfactuals—indeed, laws on the basis of which causal judgments are made in much of science—are rough-and-ready generalizations tacitly qualified by generous escape clauses ("ceteris paribus," "under normal conditions," "in the absence of interfering forces," and so on) and apparently immune to falsification by isolated negative instances. Laws of this type, sometimes called "ceteris paribus laws," seem to satisfy the usual criteria of lawlikeness: As we saw, they seem to have the power to ground counterfactuals, and their credence is enhanced as we observe more and more positive instances. Their logical form, their verification conditions, and their efficacy in explanations and predictions are not well understood, but it seems beyond question that they are the essential staple that sustains and nourishes our counterfactuals and causal discourse.[19]

Does the recognition that causal relations involving mental events can be supported by these "nonstrict," ceteris paribus laws solve the problem of mental causation? It does enable us to get around the difficulty raised by Davidsonian considerations—at least for now.

We can see, however, that the difficulty has not been fully resolved. For it may well be that these nonstrict laws are possible only if strict laws are possible and that where there are no underlying strict laws that can explain them or otherwise ground them, they remain only rough, fortuitous correlations without the power to support causal claims. It may be that their lawlike appearance is illusory and that this makes them unfit to ground causal relations. More important, as we said, the nature of these ceteris paribus laws is not well understood; though laws of this kind seem in fact used to back up causal claims, we lack a theoretical understanding of how this works.

Physical Causal Closure and the "Exclusion Argument"

Suppose, then, that we have somehow overcome the difficulties arising from the possibility that there are no mental-physical laws capable of supporting mental-physical causal relations. We are still not home free: There is another challenge to mental causation that we must confront, a challenge that is currently considered to be the gravest threat to the possibility of mental causation. The new threat arises from the principle, embraced by most physicalists, that asserts that the physical domain is *causally closed*. What does this mean? Pick any physical event—say, the decay of a uranium atom or the collision of two stars in distant space—and trace its causal ancestry or posterity as far as you would like; the principle of physical causal closure says that this will never take you outside the physical domain. Thus, no causal chain involving a physical event ever crosses the boundary of the physical into the nonphysical: If x is a physical event and y is a cause or effect of x, then y too must be a physical event.

For present purposes, it is convenient to use a somewhat weaker form of causal closure stated as follows:

> *Causal Closure of the Physical Domain*. If a physical event has a cause (occurring) at time t, it has a sufficient physical cause at t.

Notice a few things about this principle. First, it does not flatly say that a physical event can have no nonphysical cause; all it says is that in our search for its cause, we never need to look outside the physical domain. In that sense, the physical domain is causally, and hence explanatorily, self-sufficient and self-contained. Second, it does not say that every physical event has a sufficient physical cause or a physical causal explanation; in this regard, it differs from physical causal determinism, the thesis that every physical event has a sufficient physical cause. Third, the closure principle is consistent with mind-body dualism: So far as it goes, there might be a separate domain of Cartesian immaterial minds. All it requires is that there be no injection of causal influence into the physical world from outside, including Cartesian minds.

Most philosophers appear to find physical causal closure plausible; of course, anyone who considers himself or herself a physicalist of any kind must accept it. If the closure should fail to hold, there would be physical events for whose explanation we would have to look to nonphysical causal agents, like spirits or divine forces outside space-time. That is exactly the situ-

ation depicted in Descartes's interactionist dualism (chapter 2). If closure fails, theoretical physics would be in principle incompletable, a prospect that few physicists would countenance. It seems clear that research programs in physics, and the rest of the physical sciences, presuppose something like the closure principle.

It is worth noting that neither the biological domain nor the psychological domain—in fact, no domain of a special science—is causally closed: There are nonbiological events that cause biological events (for example, radiation causing cells to mutate; a volcanic eruption wiping out a whole species), and we are familiar with cases in which nonpsychological events cause psychological events (for example, purely physical stimuli causing sensations and perceptual experiences). In any case, physical causal closure gives a meaning to the widely shared view that the physical domain is an all-encompassing domain and that physics, which is the science of this domain, is our basic science. Some consider the closure principle an a posteriori truth overwhelmingly supported by the rise of modern physical science;[20] those who consider the very idea of causal interference in the physical world from some immaterial or transcendental forces incoherent might argue that the closure principle is conceptual and a priori. It is also possible to regard the principle primarily as a methodological-regulative principle that guides research and theory-building in the physical sciences.

At any rate, it is easy to see that the physical closure principle directly creates difficulties for mental causation, in particular mental-to-physical causation. Suppose that a mental event, m, causes a physical event, p. The closure principle says that there must also be a physical cause of p—an event, p^*, occurring at the same time as m, that is a sufficient cause of p. This puts us in a dilemma: Either we have to say that $m = p^*$—namely, identify the mental cause with the physical cause as a single event—or else we have to say that p has two distinct causes, m and p^*, that is, it is causally overdetermined. The first horn turns what was supposed to be a case of mental-to-physical causation into an instance of physical-to-physical causation, a result only a reductionist physicalist would welcome. Grasping the second horn of the dilemma would force us to admit that every case of mental-to-physical causation is a case of causal overdetermination, one in which a physical cause, even if the mental cause had not occurred, would have brought about the physical effect. This seems like a bizarre thing to believe, but quite apart from that, it appears to weaken the status of the mental event as a cause of the physical effect. To vindicate m as a full and genuine cause of p, we should be able to show that m can bring about p on its own, without there being a synchronous physical event that also serves as a sufficient cause of p. According to our reasoning, however, every mental event

has a physical partner that would have brought about the effect anyway, even if the mental cause were taken out of play entirely.

This thought can be developed along the following lines. Consider the following constraint:

> *Exclusion Principle*. No event has two or more distinct sufficient causes, all occurring at the same time, unless it is a genuine case of overdetermination.

Genuine overdetermination is illustrated by the "firing squad" example: Multiple bullets hit a person at the same time, and this kills the person, where a single bullet would have sufficed. A house fire is caused by a short circuit and at the same time by a lightning strike. In these cases, two or more independent causal chains converge on a single effect. Given this, the exclusion principle should look obviously, almost trivially, true.

Return now to our case of mental-to-physical causation. We begin with the assumption that there is a case in which a mental event causes a physical event:

(1) m is a cause of p.

As we saw, it follows from (1) and physical causal closure that there is also a physical event p^*, occurring at the same time as m, such that:

(2) p^* is a cause of p.

Let us suppose further that we don't want (1) to collapse into a case of physical-to-physical causation; that is, we want:

(3) $m \neq p^*$.

Suppose we assume further:

(4) This is not a case of overdetermination.

Given the closure and the exclusion principles, these four propositions put us in trouble: According to (1), (2), and (3), p has two distinct causes, m and p^*; since (4) says that this is not a case of overdetermination, the exclusion principle kicks in, saying that either m or p^* must be disqualified as a cause of p. Which one? The answer: p^* stays, m must go. The reason is simple: If we try to retain m,

the closure principle kicks in again and says that there must also be a physical cause of p—and what could this be if not p^*? Obviously, we are back at the same situation: Unless we eliminate m and keep p^*, we would be off to an infinite regress, or treading water forever in the same place. So our conclusion has to be:

(5) Hence, m is not a cause of p, and (1) is false.

The reasoning obviously generalizes to every putative case of mental causation, and it further follows:

(6) Mental events never cause physical events.

This argument is a form of the much-debated "exclusion argument," since it aims to show how a mental cause of a physical event is always excluded by a physical cause.[21] The apparent moral of the argument is that mental-to-physical causation is illusory; it never happens. This is epiphenomenalism, at least with regard to causation of physical events. It does not exclude mental events causing other mental events. But if mental events, like beliefs and intentions, never cause bodily movements, that makes agency plainly impossible, and Fodor has something to worry about: His world might be coming to an end!

That, anyway, is the way the implications of the argument are usually understood. However, that is not the only way to read the moral of the argument: If we are prepared to reject the antiphysicalist assumption (3) by embracing the mind-body identity "$m = p^*$," we can escape the epiphenomenalist consequence of the argument. If $m = p^*$, here there is only one event and hence only one cause of p, so the exclusion principle has no application and no conclusion follows to the effect that the initial supposition "m causes p" is false. The real lesson of the argument, therefore, is this: *Either accept serious physicalism, like the psychoneural identity theory, or face the specter of epiphenomenalism!*

As noted, the epiphenomenalism involved here concerns only the efficacy of mental events in the causation of physical events, not the causal power of mentality in general. However, a more radical epiphenomenalism rears its unwelcome head in the next section.

The "Supervenience Argument" and Epiphenomenalism

When you throw mind-body supervenience into the mix, an even more serious threat of epiphenomenalism arises. (The argument can be run in terms of

the idea that mental properties are "realized" by physical-neural properties rather than the premise that the former supervene on the latter.) Let us understand mind-body supervenience in the following form:

> *Mind-Body Supervenience*. When a mental property, M, is instantiated by something *x* at *t*, that is in virtue of the fact that *x* instantiates, at *t*, a physical property, P, such that anything that has P at any time necessarily has M at the same time.

So whenever you experience a headache, that is in virtue of the fact that you are in some neural state N at the time, where N is a supervenience base of headaches in the sense that anyone who is in N must be having a headache. There are no free-floating mental states; every mental state is anchored in a physical-neural base on which it supervenes.

Given mind-body supervenience, an argument can be developed that appears to have disastrous epiphenomenalist consequences.

> (1) Suppose that a mental event, an instantiation of mental property M, causes another mental property, M*, to instantiate.
>
> (2) According to mind-body supervenience, M* is instantiated on this occasion in virtue of the fact that a physical property—one of its supervenience bases—is instantiated on this occasion. Call this physical base P*.
>
> (3) Now ask: Why is M* instantiated on this occasion? What is responsible for the fact that M* occurs on this occasion? There appear to be two presumptive answers: (i) because an instance of M caused M* to instantiate (our original supposition), and (ii) because a supervenience base, P*, of M*, was instantiated on this occasion.

Now, there appears to be strong reason to think that (ii) trumps (i): If its supervenience base P* occurs, M* must occur, no matter what preceded M*'s occurrence—that is, *as long as P* is there, M* is guaranteed to be there even if its supposed cause M did not occur*. This undermines M's claim to have brought about this instance of M*; it seems that P* must take the primary credit for bringing about M* on this occasion. Is there a way of reconciling M's claim to have caused M* to instantiate and P*'s claim to be M*'s supervenience base on this occasion?

(1) The two claims (i) and (ii) can be reconciled if we are willing to accept: M caused M* to instantiate *by causing M*'s supervenience base P* to instantiate*. This seems like the only way to harmonize the two claims.

In general, it seems like a plausible principle to say that in order to cause, or causally affect, a supervenient property, you must cause, or tinker with, its supervenience base. If you are not happy with a painting you have just finished and want to improve it, there is no way you could alter the aesthetic qualities of the painting (for example, make it more expressive, more dramatic, and less sentimental) except by altering the physical properties on which the aesthetic properties supervene. You must bring out your brushes and oils and do physical work on the canvas. That is the only way. You take aspirin to relieve your headache because you hope that ingesting aspirin will bring about physicochemical changes in the neural state on which your headache supervenes.

(2) Hence, M causes P*. This is an instance of mental-to-physical causation.

If this argument is correct, it shows that, given mind-body supervenience, mental-to-mental causation (an instance of M causing M* to instantiate) leads inevitably to mental-to-physical causation. This argument, which may be called the "supervenience argument," shows that mental-to-mental causation is possible only if mental-to-physical causation is possible.

But see where the two arguments, the exclusion argument and the supervenience argument, lead us. According to the supervenience argument, mental-to-mental causation is possible only if mental-to-physical causation is possible. But the exclusion argument says that mental-to-physical causation is not possible. So it follows that neither mental-to-mental causation nor mental-to-physical causation is possible. This goes beyond the epiphenomenalism of mental-to-physical causation; the two arguments together purport to show that mental events have no causal efficacy at all, no power to cause any event, mental or physical. This is radical epiphenomenalism.

It is important to keep in mind that all this holds on the assumption that we do not choose the option of reductionist physicalism; that is, if we reject the premise "$m \neq p^*$" of the exclusion argument, thereby accepting the psychoneural identity "$m = p$," we can avoid the epiphenomenalist conclusion. So the upshot of these two arguments is this: If you want to avoid radical epiphenomenalism, you must be prepared to embrace reductionist physicalism—that is, you must choose between an extreme form of epiphenomenalism and reductionism.

Neither option is palatable. To most of us, epiphenomenalism seems just false, or even incoherent (recall Fodor's lament). And reductionist physicalism does not seem much better: If we save mental causation by reducing mentality to mere patterns of electrochemical activity in the brain, have we really saved mentality as something special and distinctive? Moreover, what if the mental is not reducible to the physical? Aren't we then stuck with epiphenomenalism whether we like it or not? This is the conundrum of mental causation.

The general moral of our discussion seems to be this: If anything is to have causal powers and enter into causal relations with anything else, it must be part of the physical domain. This conclusion complements, and strengthens, what we learned about the problem of mental causation for Descartes's immaterial minds (chapter 2).

Further Issues: The Extrinsicness of Mental States

Computers compute with 0s and 1s. Suppose you have a computer running a certain program, say, a program that monitors the inventory of a supermarket. Given a string of 0s and 1s as input (a can of Campbell's tomato soup has just been scanned at a checkout station), the computer goes through a series of computations and emits an output (the count of Campbell's tomato soup in stock has been adjusted, and so on). Thus, the input string of 0s and 1s represents a can of Campbell's tomato soup being sold, and the output string of 0s and 1s represents the amount of Campbell's tomato soup still in stock. When the manager checks the computer for a report on the available stock of Campbell's tomato soup, the computer "reports" that the present stock is such and such, and it does so *because* "it has been told" (by the checkout scanners) that twenty-five cans have been sold so far today. And this "because" is naturally understood as signifying a causal relation.

But we know that it makes no difference to the computer what the strings of 0s and 1s *mean* or *represent*. If the input string had meant the direction and speed of wind at the local airport or the identification code of an employee, or even if it had meant nothing at all, the computer would have gone through exactly the same computation and produced the same output string. In this case, the output string too would have meant something else, but what is clear is that the "meanings," or "representational contents," of these 0s and 1s are in the eye of the computer programmer or user, not something that is involved in the computational process. Give the computer the same string of 0s and 1s as input, and it will go through the same computation every time and give you the

same output. The "semantics" of these strings is irrelevant to computation; what matters is their shape—that is, their syntax. The computer is a "syntactic engine"; it is driven by the shapes of symbols, not their meanings.

According to an influential view of psychology known as computationalism (or the computational theory of mind), cognitive mental processes are best viewed as computational processes on mental representations (chapter 5). According to it, constructing a psychological theory is like writing a computer program; such a theory will specify, for each input (say, retinal stimulation), the computational process that a cognizer will undergo to produce an output (say, the visual detection of an edge). But what the considerations of the preceding paragraph seem to show is that, on the computational view of psychology, the meanings, or contents, of internal representations make no difference to psychological processes. Suppose a certain internal representation, *i*, represents the state of affairs S (say, that there are horses in the field); having S as its representational content, or meaning, is the semantics of *i*. But if we suppose, as is often done on the computational model, that internal representations form a language-like system (the "language of thought"), *i* must also have a syntax, or formal grammatical structure. So if our considerations are right, it is the syntax of *i*, not its semantics, that determines the course of the computational process starting with *i*. The fact that *i* means that there are horses in the field rather than, say, that there are lions in the field, is of no causal relevance to what other representations issue from *i*. The computational process that *i* initiates will be wholly determined by *i*'s syntactic shape. But doesn't this mean that the *contents* of our beliefs and desires and of other propositional attitudes have no causal relevance for psychological processes?

The point actually is independent of computationalism and can be seen to arise for any broadly physicalist view of mentality. Assume that beliefs and desires and other intentional states are neural states. Each such state, in addition to being a neural state with biological-physical properties, has a specific content (for example, that water is wet, or that the Obamas have a home in Chicago). That a given state has the content it has is a *relational*, or *extrinsic*, *property* of that state, for the fact that your belief is about water, or about the Obamas, is in part determined by your causal-historical associations with water and the Obamas (see chapter 8). Let us consider what this means and why it is so.

Suppose there is in some remote region of this universe another planet, "Twin Earth," that is exactly like our Earth, except for the following fact: On Twin Earth, there is no water, that is, no H_2O, but an observably indistinguishable chemical substance, XYZ, fills the lakes and oceans there, comes out of the tap in Twin Earth homes, and so on. Each of us has a doppelganger there who

is an exact molecular duplicate of us. (Let us ignore the inconvenient fact that your twin has XYZ molecules in her body where you have H_2O molecules.) On Twin Earth, people speak Twin English, which is just like English, except for the fact that their word "water" refers to XYZ, not water, and when they utter sentences containing the expression "water," they are talking about XYZ, not water. Thus, Twin Earth people have thoughts about XYZ, where we have thoughts about water, and when you believe that water is wet, your doppelganger on Twin Earth has the belief that XYZ is wet, even though you and she are molecule-for-molecule duplicates. And when you think that Obama is from Chicago, your twin thinks that the Twin Earth Obama (he is the forty-fourth president of the Twin Earth United States) is from Twin Earth Chicago. And so on. The differences in Earth and Twin Earth belief contents (and contents of other intentional states) are due not to internal physical or mental differences in the believers but to the differences in the environments in which the believers are embedded (see the discussion of "wide content" in chapter 8). Contents, therefore, are extrinsic, not intrinsic; they depend on your causal history and your relationships to the objects and events in your surroundings. States that have the same intrinsic properties—the same neural-physical properties—may have different contents if they are embedded in different environments. Further, an identical internal state that lacks an appropriate relationship to the external world may have no representational content at all.

But isn't it plausible to suppose that behavior causation is "local" and depends only on the intrinsic neural-physical properties of these states, not their extrinsic relational properties? Isn't it plausible to suppose that someone whose momentary neural-physical state is exactly identical with yours will behave just the way you do—say, raise the right hand—regardless of whether her brain state has the same content as yours? This raises doubts about the causal relevance of contents because the properties of our mental states implicated in behavior causation are plausibly expected to be intrinsic. What causes your behavior, we feel, must be *local—in you*, *here and now*; after all, the behavior it is supposed to cause is here and now. But contents of mental states are relational and extrinsic; they depend on what is out there in the world outside you, or on what occurred in the past and is no longer here. To summarize, contents do not supervene on the intrinsic properties of the states that carry them; on the other hand, we expect behavior causation to be local and depend only on intrinsic properties of the behaving organism. This, then, is yet another problem of mental causation. It challenges us to answer the following question: How can intentional mental states, like beliefs and desires, be efficacious in behavior causation in virtue of their contents?

Various attempts have been made to reconcile the extrinsicness of contents with their causal efficacy, but we do not as yet have a fully satisfactory account. The problem has turned out to be a highly complex one involving many issues in metaphysics, philosophy of language, and philosophy of science.[22]

For Further Reading

Donald Davidson's "Mental Events" is the primary source of anomalous monism. On the problem of mental causation associated with anomalous monism, see Ernest Sosa, "Mind-Body Interaction and Supervenient Causation," and Louise Antony, "Anomalous Monism and the Problem of Explanatory Force." Davidson responds in "Thinking Causes," which appears in *Mental Causation*, edited by John Heil and Alfred Mele. This volume also contains rejoinders to Davidson by Kim, Sosa, and Brian McLaughlin, as well as a number of other papers on mental causation.

For counterfactual-based accounts of mental causation, see Ernest LePore and Barry Loewer, "Mind Matters," and Terence Horgan, "Mental Quausation." On functionalism and mental causation, see Ned Block, "Can the Mind Change the World?" and Brian McLaughlin, "Is Role-Functionalism Committed to Epiphenomenalism?"

Journal of Consciousness Studies, vol. 13, no. 1–2, edited by Michael Pauen, Alexander Staudacher, and Sven Walter, is a special issue on epiphenomenalism and contains many interesting papers on the topic.

For issues related to the causal role of extrinsic mental states, see Fred Dretske, "Minds, Machines, and Money: What Really Explains Behavior," and Tim Crane, "The Causal Efficacy of Content: A Functionalist Theory." Many of the issues in this area are discussed in Lynne Rudder Baker, *Explaining Attitudes*; Dretske, *Explaining Behavior*; Pierre Jacob, *What Minds Can Do*. Stephen Yablo's "Wide Causation" is interesting but difficult and challenging.

On the principle of physical causal closure, see David Papineau's "The Rise of Physicalism" and "The Causal Closure of the Physical and Naturalism." For a different perspective, see E.J. Lowe, "Physical Causal Closure and the Invisibility of Mental Causation" and "Non-Cartesian Substance Dualism and the Problem of Mental Causation."

On the exclusion and supervenience arguments, see Jaegwon Kim, *Mind in a Physical World* and *Physicalism, or Something Near Enough*, chapter 2. Many interesting papers on issues on these and related topics are found in *Physicalism and Mental Causation*, edited by Sven Walter and Heinz-Dieter Heckmann. Recommended also are Stephen Yablo, "Mental Causation"; Karen

Bennett, "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It" and "Exclusion Again"; John Gibbons, "Mental Causation Without Downward Causation." For an interesting and wide-ranging discussion of the exclusion principle and related issues, see Christian List and Peter Menzies, "Nonreductive Physicalism and the Limits of the Exclusion Principle."

Some philosophers advocate the "trope" theory as basic ontology, in order to get around some of the difficulties with mental causation. A good example is "The Metaphysics of Mental Causation" by Cynthia Macdonald and Graham Macdonald.

Karen Bennett's "Mental Causation" is a balanced and accessible overview and discussion of mental causation.

Notes

1. Marcel Proust, *Remembrance of Things Past*, vol. 1, pp. 48–51.

2. Some philosophers insert another step between beliefs-desires and actions, by taking beliefs-desires to lead to the formation of *intentions* and *decisions*, which in turn lead to actions. What has been described is the influential causal theory of action, which is widely, but far from universally, accepted. Details concerning action, agency, and action explanation are discussed in a subfield of philosophy called action theory, or the philosophy of action.

3. Whether explanations appealing to emotions presuppose belief-desire explanations is a controversial issue. For discussion, see Michael Smith, "The Possibility of Philosophy of Action."

4. Donald Davidson, "Actions, Reasons, and Causes." For noncausal approaches, see Carl Ginet, *On Action*, and Frederick Stoutland, "Real Reasons."

5. Thomas H. Huxley, "On the Hypothesis That Animals Are Automata, and Its History," *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers, pp. 29–30.

6. Huxley advances his epiphenomenalism in regard to consciousness; it isn't clear what his views are about the causal status of mental states like beliefs and desires. Does the French sergeant perform actions? Does he have beliefs and desires?

7. Samuel Alexander, *Space, Time, and Deity*, vol. 2, p. 8.

8. Jerry A. Fodor, "Making Mind Matter More," in Fodor, *A Theory of Content and Other Essays*, p. 156.

9. More precisely, Davidson's claim is that there are no "strict" laws connecting psychological and physical phenomena. There are some questions about

what the strictness of laws amounts to; for our present purposes, it is sufficient to understand "strict" as "exceptionless." See Davidson's "Mental Events."

10. See Donald Davidson's "Mental Events." For an interpretive reconstruction of Davidson's argument, see Jaegwon Kim, "Psychophysical Laws."

11. This is a form of what is called "the principle of charity"; Davidson also requires that an interpretation of a person's belief system make her beliefs come out largely *true*. See the discussion of interpretation theory in chapter 8.

12. In "Mental Events," Davidson defends the stronger thesis that there are no laws at all about mental phenomena, whether psychophysical or purely psychological; his view is that laws (or "strict laws") can be found only in basic physics (see "Thinking Causes"). A sharp-eyed reader will have noticed that Davidson's argument requires this stronger thesis, since the argument as it stands leaves open the possibility that the two causally connected events, *m* and *p*, instantiate a purely psychological law, from which it would follow that *p* is a mental event. If, as Davidson believes, "strict" laws are found only in physics, his conclusion can be strengthened: Any event (of any kind) that causes, or is caused by, another event (of any kind) is a physical event. For a defense of the thesis that there are no laws at all about psychological phenomena, see Jaegwon Kim, "Why There Are No Laws in the Special Sciences: Three Arguments."

13. Brian McLaughlin calls it "type epiphenomenalism" in his "Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical." Several philosophers independently raised these epiphenomenalist difficulties for anomalous monism; Frederick Stoutland was probably the first to do so, in his "Oblique Causation and Reasons for Action."

14. This is not quite complete. The counterfactual analysis of causation only requires that there be a chain of these "counterfactual dependencies" connecting cause and effect. But this and other refinements do not affect the discussion to follow. David Lewis's "Causation" is the first full counterfactual analysis of causation.

15. Of late, the possible-world semantics has been dominant for counterfactuals; the first approach has virtually disappeared from the scene.

16. See Ernest Nagel, *The Structure of Science*, chapter 4.

17. For a detailed development of this approach, see David Lewis, *Counterfactuals*. Lewis's account does not require that there be "the closest" P-world; there could be ties.

18. These worlds are very much underdescribed, of course; we are assuming that the worlds are roughly the same in other respects.

19. Later in his career, Davidson too came to accept nonstrict laws as capable of grounding causal relations; see his "Thinking Causes." But this may very well undermine his argument for anomalous monism.

20. See David Papineau, "The Rise of Physicalism."

21. For more detail, see Jaegwon Kim, *Physicalism, or Something Near Enough*, chapter 2.

22. The inability to reach a satisfactory solution to this problem can add fuel to the eliminativist argument on content-carrying mental states, along the lines urged by Paul Churchland in "Eliminative Materialism and the Propositional Attitudes." If contents are causally inefficacious, how can they play a role in causal-explanatory accounts of human behavior? And if they can have no such role, why should we bother with them, whether in commonsense psychology or the science of human behavior?

CHAPTER 8

Mental Content

You hope that it will be warmer tomorrow, and I believe that it will be. But Mary doubts it and hopes that she is right. Here we have various "intentional" (or "content-bearing" or "content-carrying") states: your *hoping* that it will be warmer tomorrow, my *believing*, and Mary's *doubting*, that it will be so. All of these states, though they are states of different persons and involve different *attitudes* (believing, hoping, and doubting), have the same *content*: the proposition that it will be warmer tomorrow, expressed by the embedded sentence "it will be warmer tomorrow." This content *represents* a certain state of affairs, its being warmer tomorrow. Different subjects can adopt the same intentional attitude toward it, and the same subject can have different attitudes toward it (for example, you believe it and are pleased about it; later you come to disbelieve it).

But how do these intentional states, or propositional attitudes, come to have the content they have and represent the state of affairs they represent? More specifically, what makes it the case that your hope and my belief have the same content? There is a simple, and not wholly uninformative, answer: Because they each have the content expressed by the same content sentence "it will be warmer tomorrow." But then a more substantive question awaits us: What is it about your hope and my belief that makes it the case that the same sentence can capture their content? We do not expect it to be a brute fact about these mental states that they have the content they have or that they share the same content; there must be an explanation. These are the basic questions about mental content.

The questions can be raised another way. It is not just persons who have mental states with content. All sorts of animals perceive their surroundings through their perceptual systems, process information gained thereby, and use it in coping with things and events around them. We humans do this in our own distinctive ways, though perhaps not in ways that are fundamentally different

from those of other higher species of animals. It seems, then, that certain physical-biological states of organisms, presumably states of their brains or nervous systems, can carry information about their surroundings, representing them as being this way or that way (for example, here is a red apple, or a large, brown, bear-shaped hulk is approaching from the left), and that processing and using these representations in appropriate ways is highly important to their surviving and flourishing in their environments. These physical-biological states have representational content—they are *about* things, inside or outside an organism, and *represent them as being a certain way*. In a word, these states have *meanings*: A neural state that represents a bear as approaching *means* that a bear is approaching. But how do neural-physical states come to have meanings—and come to have the particular meanings that they have? Just what is it about a configuration of nerve fibers or a pattern of their activation that makes it carry the content "there is a red apple on the table" rather than, say, "there are cows in Canada," or perhaps nothing at all?

This question about the nature of mental content has a companion question, a question about how contents are *attributed* to the mental states of persons and other intentional systems. We routinely ascribe states with content to persons, animals, and even some nonbiological systems. If we had no such practice—if we were to stop attributing to people around us beliefs, desires, emotions, and the like—our communal life would surely suffer a massive collapse. There would be little understanding or anticipating of what other people will do, and this would seriously undermine interpersonal interactions. Moreover, it is by attributing these states to ourselves that we come to understand ourselves as cognizers and agents. A capacity for self-attribution of beliefs, desires, intentions, and the rest is arguably a precondition of personhood. Moreover, we often attribute such states to nonhuman animals and sometimes even to purely mechanical or electronic systems. (Even such humble devices as supermarket doors are said to "see that a customer is approaching.") What makes it possible for us to attribute content-carrying states to persons and other organisms? What procedures and principles do we follow when we do this? According to some philosophers, the two questions, one about the nature of mental content and the other about its attribution, are intimately connected.

Interpretation Theory

Suppose you are a field anthropologist-linguist visiting a tribe of people never before visited by an outsider. Your project is to find out what these people believe, remember, desire, fear, hope, and so on, and to be able to understand their

speech. That is, your project is to map their "notional world" and develop a grammar and dictionary for their language. So your job involves two tasks: first, interpreting their minds, to find out what they believe, desire, and so on; and, second, interpreting their speech, to determine what their utterances mean. This is the project of "radical interpretation": You are to construct an interpretation of the natives' speech and their minds from scratch, based on your observation of their behavior and their environment, without the aid of a native translator-informant or a dictionary. (This is what makes it "radical" interpretation.)[1]

Brief reflection shows that the twin tasks are interconnected and interdependent. In particular, belief, among all mental states, can be seen to hold the key to radical interpretation: It is the crucial link between a speaker's utterances and their meanings. If a native speaker sincerely asserts sentence S (or more broadly, "holds S true," as Donald Davidson says) and S means that there goes a rabbit, then the speaker believes that there goes a rabbit, and in asserting S she expresses her belief that there goes a rabbit. Conversely, if the speaker believes that there goes a rabbit and uses sentence S to express this belief, then S means that there goes a rabbit. If you knew how to interpret the natives' speech, it would be a simple matter to find out what they believe: All you would need to do is observe their speech behavior—their assertions, denials, and so on. Similarly, if you had knowledge of what belief a native is expressing by uttering S on a given occasion, you know what S, as a sentence of her language, means. When you begin, you have knowledge of neither her beliefs nor her meanings, and your project is to secure them both through your observation of how she behaves in her environment. There are, then, three variables involved: behavior, belief, and meaning. Through observation, you have access to one of them, behavior. Your task is to solve for the two unknowns, belief and meaning. How is this possible? Where do you start?

Karl is one of the subjects you are trying to interpret. Suppose you observe that Karl affirmatively utters, or holds true,[2] the sentence "Es regnet" when, and only when, it is raining in his vicinity. (This is highly idealized, but the main point should apply, with suitable provisos, to real-life situations.) You observe a similar behavior pattern in many others in Karl's speech community, and you are led to posit the following proposition:

> (R) Speakers of language L (Karl's language) utter "Es regnet" at time *t* if and only if it is raining at *t* in their vicinity.

So we are taking (R) to be something we can empirically establish by observing the behavior, in particular, speech behavior, of our subjects in the context

of what is happening in their immediate environment. Assuming, then, that we have (R) in hand, it would be natural to entertain the following two hypotheses:

> (S) In language L, "Es regnet" means that it is raining (in the speaker's vicinity).
>
> (M) When speakers of L utter "Es regnet," this indicates that they believe that it is raining (in their vicinity) and they use "Es regnet" to express this belief.

In this way you get your first toehold in the language and minds of the natives, and something like this seems like the only way.

These hypotheses, (S) and (M), are natural and plausible. But what makes them so? What sanctions the move from (R) to (S) and (M)? When you observe Karl uttering the words "Es regnet," you see yourself that it is raining out there. You have determined observationally that Karl is expressing a belief about the current condition of the weather. This assumption is reinforced when you observe him, and others in his speech community, do this time after time. But what belief is Karl expressing when he makes this utterance? What is the content of the belief that Karl expresses when he says "Es regnet"? Answering this question is the crux of the interpretive project. The obvious answer seems to be that Karl's belief has the content "it is raining." But why? Why not the belief with the content "it is a sunny day" or "it is snowing"? What are the tacit principles that help to rule out these possibilities?

You attribute the content "it is raining" to Karl's belief *because you assume that his belief is true*. You know that his belief is about the weather outside, and you see that it is raining. What you need, and all you need, to get to the conclusion that his belief has the content "it is raining" is the further premise that his belief is true. In general, then, what you need is the famous "charity principle":

> *Principle of Charity*. Speakers' beliefs are by and large true. (Moreover, they are largely correct in making inferences and rational in forming expectations and making decisions.)[3]

With this principle in hand, we can make sense of the transition from (R) to (S) and (M) in the following way:

> In uttering "Es regnet," Karl is expressing a belief about the current weather condition in his vicinity, and we assume, by the charity principle, that this belief is true. The current weather condition is that it is raining. So Karl's belief has the content that it is raining, and he is using the sentence "Es regnet" to express this belief (M), whence it further follows that "Es regnet" means that it is raining (S).

We do not attribute the content "it is clear and sunny" or "it is snowing" because that would make Karl's and his friends' beliefs about whether it is raining around them almost invariably, and unaccountably, false. There is no logical contradiction in the idea that a group of speakers are almost always wrong about rains in their vicinity, but it is not something that can be taken seriously. We would have to posit serious, and unexplainable, cognitive deficits in Karl and his friends, and this is not a reasonable possibility. For one thing, they seem able to cope with their surroundings, including good and bad weather, as well as we do.

Clearly, the same points apply to interpreting utterances about colors, shapes, and other observable properties of objects and events around Karl. When Karl and his friends invariably respond with "Rot" when we show them cherries, ripe tomatoes, and McIntosh apples and withhold it when they are shown lemons, eggplants, and snowballs, it would make no sense to speculate that "rot" might mean *green*, that Karl and his friends systematically misperceive colors, and that in consequence they have massively erroneous beliefs about the colors of objects around them. The only plausible thing to say is that "rot" means *red* in Karl's language and that Karl is expressing the (true) belief that the apple held in front of him is red. All this is not to say that our speakers never have false beliefs about colors or about anything else; they may have them in huge numbers. But unless we assume that their beliefs, especially those about the manifestly observable properties of things and events around them, are largely correct, we have no hope of gaining entry into their notional world.

So what happens is that we interpret the speakers in such a way as to credit them with beliefs that are by and large true and coherent. But since *we* are doing the interpreting, this in effect means *true and coherent by our light*. Under our interpretation, therefore, our subjects come out with *beliefs that are largely in agreement with our own*. The attribution of a system of beliefs and other intentional states is essential to the understanding of other people, of what they say and do. From all this an interesting conclusion follows: We can

interpret and understand only those people whose belief systems are largely like our own.

The charity principle therefore rules out, a priori, interpretations that attribute to our subjects beliefs that are mostly false or incoherent; any interpretive scheme according to which our subjects' beliefs are massively false or manifestly inconsistent (for example, they come out believing that there are round squares) cannot, for that very reason, be a correct interpretation. Further, we can think of a generalized charity principle that enjoins us to interpret all of our subjects' intentional states, including desires, aversions, hopes, fears, and the rest, in a way that renders them maximally coherent and intelligible among themselves and in relation to the subjects' actions and behaviors.

But we should note the following important point: There is no reason to think that in any interpretive project there is a single unique interpretation that best meets this requirement. This is evident when we reflect on the fact that the charity principle requires only that the entire *system* of beliefs attributed to a subject be by and large true but it does not tell us which of her beliefs must come out true. In practice as well as in theory, there are likely to be ties, or unstable near-ties, among possible interpretations: That is, we are likely to end up with more than one maximally true, coherent, and rational scheme of interpretation that can explain all the observational data. (This phenomenon is called "indeterminacy of interpretation.") We can appreciate such a possibility when we note that our criteria of coherence and rationality are bound to be somewhat vague and imprecise (in fact, this is probably necessary to ensure their flexible application to a wide and unpredictable range of situations) and that their applications to specific situations are likely to be fraught with ambiguities. At any rate, it is easy to see how interpretational indeterminacy can arise by considering a simple example.

We observe Karl gorging on raw spinach leaves. Why is he doing that? We can see that there are indefinitely many belief-desire pairs that we could attribute to Karl that would explain why he is eating raw spinach. The following are only some of the possibilities:

Karl believes that eating raw spinach will improve his stamina, and he wants to improve his stamina.

Karl believes that eating raw spinach will help him get rid of his bad breath, and he has been very self-conscious about his breath.

Karl believes that eating raw spinach will please his mother, and he will do anything to make her happy.

> Karl believes that eating raw spinach will annoy his mother, and he will go to any length to annoy her.

You get the idea: This can go on without end. We can expect many of these potential explanations to be excluded by further observation of Karl's behavior and by consideration of coherence with other beliefs and desires that we want to attribute to him. But it is difficult to imagine that this will eliminate all but one of the indefinitely many possible belief-desire pairs that can explain Karl's spinach eating. Moreover, it is likely that any one of these pairs could be protected no matter what if we were willing to make drastic enough adjustments elsewhere in Karl's total system of beliefs, desires, and other mental states.

Suppose, then, that there are two interpretive schemes of Karl's mental states that, as far as we can determine, satisfy the charity principle to the same degree and work equally well in explaining his behavior. Suppose further that one of these systems attributes to Karl the belief that eating raw spinach is good for one's stamina, and the second instead attributes to him the belief that eating spinach will please his mother. As far as interpretation theory goes, the schemes are in a tie, and neither could be pronounced to be superior to the other. But what is the fact of the matter concerning Karl's belief system? Does he or doesn't he believe that eating raw spinach improves stamina?

There are two possible approaches we could take in response to these questions. The first is to take interpretation as the rock-bottom foundation of content-carrying mental states by embracing a principle like this:

> For S to have the belief that *p* is for that belief to be part of the best (most coherent, maximally true, and so on) interpretive scheme of S's total system of propositional attitudes (including beliefs, desires, and the rest). There is no further fact of the matter about whether S believes that *p*.

It will be natural to generalize this principle so that it applies to all propositional attitudes, not just beliefs. On this principle, then, interpretation is *constitutive* of intentionality; it is what ultimately determines whether any supposed belief exists.[4] Interpretation is not merely a procedure for finding out what Karl believes. This constitutive view of interpretation, when combined with the indeterminacy of interpretation, can be seen to have some apparently puzzling consequences. Suppose that several interpretive schemes are tied for first and the belief that *p* is an element of some but not all of these schemes. In such

a case we would have to conclude that there is no fact of the matter about whether Karl has this belief. Whether Karl believes that *p* therefore is a question without a determinate answer. To be sure, the question about this particular bit of belief may be settled by further observation of Karl; however, indeterminacies are almost certain to remain even when all the observations are in. (Surely, at some point after Karl's death, there is nothing further to observe that will be relevant!) Some will see in this kind of position a form of *content irrealism*. If beliefs are among the objectively existing entities of the world, either Karl believes that raw spinach is good for his stamina or he does not. There must be a fact about the existence of this belief, independent of any interpretive scheme that someone might construct for Karl. So if the existence of beliefs is genuinely indeterminate, we would have to conclude, it seems, that beliefs are not part of objective reality. Evidently, the same conclusion would apply to all intentional states.[5]

An alternative line of consideration can lead to *content relativism* rather than content irrealism: Instead of accepting the indeterminacy of belief, we might hold that whether a given belief exists is *relative to a scheme of interpretation*. It is not a question that can be answered absolutely, independently of a choice of an interpretive scheme. Whether Karl has that particular belief depends on the interpretive theory relative to which we view Karl's belief system. But a relativism of this kind is not free from difficulties either. What is it for a belief to "exist relative to a scheme" to begin with? Is it anything more than "the scheme attributes the belief to Karl"? If so, shouldn't we ask the further question whether what the scheme says is *correct*? But this takes us right back to the nonrelativized notion of belief existence. Moreover, is all existence relative to some scheme or other, or is it just the existence of belief and other propositional attitudes that is relative in this way? Either way, many more questions and puzzles await us.

There is a further point to think about: Interpretation involves an interpreter, and the interpreter herself is an intentional system, a person with beliefs, desires, and so forth. How do we account for *her* beliefs and desires—*how do her intentional states get their contents*? And when she tries to maximize agreement between her beliefs and her subject's beliefs, how does she know what she believes? That is, *how is self-interpretation possible*? Don't we need an account of how we can know the contents of our own beliefs and desires? Do we just look inward, and are they just there for us to "see"? Or do we need to be interpreted by a third person if we are to have beliefs and meaningful speech? It is clear that the interpretation approach to mental content must, on pain of circularity, confront the issue of self-interpretation.

All this may lead you to reject both the constitutive and the relativist views of interpretation and pull you toward a realist position about intentional states, which insists that there is a fact of the matter about the existence of Karl's belief about spinach that is independent of any interpretive schemes. If Karl is a real and genuine believer, there must be a determinate answer to the question whether he has this belief. Whether someone happens to be interpreting Karl, or what any interpretive scheme says about Karl's belief system, should be entirely irrelevant to that question. This is content realism, a position that views interpretation only as a way of finding out something about Karl's belief system, not as constitutive of it. Interpretation therefore is given only an epistemological function, that of ascertaining what intentional states a given subject has; it does not have the ontological role of grounding their existence.

You may find content realism appealing. If so, there is more work to do; you must provide an alternative realist account of what constitutes the content of intentional states. It is only if you take the constitutive view of interpretation that interpretation theory gives you a solution to the problem of mental content—that is, an answer to the question "How does a belief get to have the content it has?"

The Causal-Correlational Approach: Informational Semantics

A fly flits across a frog's visual field, and the frog's tongue darts out, snaring the fly. The content of the frog's visual perception is a moving fly (which is a complicated way of saying that the frog sees a moving fly). Suppose now that in a world pretty much like our own (this could be some remote region of this world), frogs that are like our frogs exist but there are no flies. Instead there are "schmies," very small lizards roughly the size, shape, and color of earthly flies, and they fly around just the way our flies do and are found in the kind of habitat that our flies inhabit. In that world frogs feed on schmies, not flies. Now, in this other world, a schmy flits across a frog's visual field, and the frog flicks out its tongue and catches it. What is the content of this frog's visual perception? What does the frog's visual percept represent? The answer: a moving schmy.

From the frogs' "internal," or "subjective," perspectives, there is no difference, we may suppose, between our frog's perceptual state and the other-worldly frog's perceptual state: Both register a black speck flitting across the visual field. However, we attribute different contents to them, and the difference lies outside the frogs' perceptual systems; it is a difference in the kind of object that stands in a certain relationship to the perceptual states of the frogs. It is not only that in

these particular instances a fly caused the perceptual state of our frog and a schmy caused a corresponding state in the other-worldly frog; there is also a more general fact, namely, that the habitat of earthly frogs includes flies, not schmies, and it is flies, not schmies, with which they are in daily perceptual and other causal contact. The converse is the case with other-worldly frogs and schmies. Our frogs' perceptual episodes involving a flitting black speck *indicate*, or *mean*, the presence of a fly; qualitatively indistinguishable perceptual episodes in other-worldly frogs *indicate* the presence of a schmy.

Consider a mercury thermometer: The height of the column of mercury indicates the ambient air temperature. When the thermometer registers 32°C, we say, "The thermometer says that the temperature is 32°C"; we also say that the current state of the thermometer carries the information that the air temperature is 32°C. Why? Because there is a lawful correlation—in fact, a causal connection—between the state of the thermometer (that is, the height of its mercury column) and air temperature. It is for that reason that the device is a thermometer, something that carries *information* about ambient temperature.

Suppose that under normal conditions a certain state of an organism covaries regularly and reliably with the presence of a horse. That is, this state occurs in you when, and only when, a horse is present in your vicinity (and you are awake and alert, sufficient illumination is present, you are appropriately oriented in relation to the horse, and so on). The occurrence of this state, then, can serve as an *indicator*[6] of the presence of a horse; it carries the information "horse" (or "a horse is out there"). And it seems appropriate to say that this state *indicates* or *represents* the presence of a horse and has it as its content. The suggestion is that something like this account works for intentional content in general, and this is the basic idea of the causal-correlational approach. (The term "causal" is used because on some accounts based on this approach, the presence of horses is supposed to cause the internal "horse-indicator" state.)

The strategy seems to work well with contents of perceptual states, as we saw in the fly-schmy case. I perceive red, and my perceptual state has "red" as its content because I am having the kind of perceptual experience typically correlated with—in fact, caused by—the presence of a red object. Whether I perceive red or green has little to do with the intrinsic experienced qualities of which I am conscious; rather, it depends essentially on the properties of the objects with which I am in causal-correlational relations. Those internal states that are typically caused by red objects, or that lawfully correlate with the presence of red objects nearby, have the content "red" for that very reason, not because of any of their intrinsic properties. Two thermometers of very different construction—

say, a mercury thermometer and a gas thermometer—both represent the temperature to be 30°C in spite of the fact that the internal states of the two thermometers that covary with temperature—the height of a column of mercury in the first and the pressure of a gas in the second—are different. In a similar way, two creatures, belonging to physiologically quite diverse species, can both have the belief that there are red fruits on the tree. The causal-correlational approach to content, also called informational semantics, has been influential; it explains mental content in a naturalistic way and seems considerably simpler than the interpretational approach considered earlier.

How well does this approach work with intentional states in general? We may consider a simple version of this approach, perhaps something like this:[7]

> (C) Subject S has the belief with content *p* (that is, S believes that *p*) just in case, under optimal conditions, S has this belief (as an occurrent belief)[8] if and only if *p* obtains.

To make (C) at all viable, we should restrict it to cases of "observational beliefs"—beliefs about matters that are perceptually observable to S. For (C) is obviously implausible when applied to beliefs like the belief that God exists or that light travels at a finite velocity and beliefs about abstract matters (say, the belief that there is no largest prime number). It is much more plausible for observational beliefs like the belief that there are red flowers on my desk or that there are horses in the field. The proviso "under optimal conditions" is included since for the state of affairs *p* (for example, the presence of horses) to correlate with, or cause, subject S's belief that *p*, favorable perceptual conditions must obtain, such as that S's perceptual systems are functioning properly, the illumination is adequate, S's attention is not seriously distracted, and so on.

Although there seem to be some serious difficulties that (C) has to overcome, remember that (C) is only a rough-and-ready first pass, and none of the objections enumerated here need be taken as a disabling blow to the general approach.

1. The belief that there are horses in the field correlates reliably, let us suppose, with the presence of horses in the field. But it also correlates reliably with the presence of horse genes in the field (since the latter correlate reliably with the presence of horses). According to (C), someone observing horses in the field should have the belief that there are horse genes in the field. But this surely is wrong.

Moreover, the belief that there are horses in the field also correlates with the presence of undetached horse parts. But again, the observer does not have the belief that there undetached horse parts in the field. The general problem, then, is that an account like (C) cannot differentiate between belief with *p* as its content and belief with *q* as its content if *p* and *q* reliably correlate with each other. For any two correlated states of affairs *p* and *q*, (C) entails that one believes that *p* if and only if one believes that *q*, which evidently is incorrect. Restricting (C) to observational beliefs can relieve some of this problem, however.

2. Belief is *holistic* in the sense that what you believe is shaped, often crucially, by what else you believe. When you observe horselike shapes in the field, you are not likely to believe that there are horses in the field if you have read in the papers that many cardboard horses have been put up for a children's fair, or if you believe you are hallucinating, and so on. Correlational accounts make beliefs basically atomistic, at least for observational beliefs, but even our observational beliefs are constrained by other beliefs we hold, and the correlational approach as it stands is not sensitive to this aspect of belief content.
3. The belief that there are horses in the field is caused not only by horses in the field but also by cows and moose at dusk, cardboard horses at a distance, robot horses, and so on. In fact, this belief correlates more reliably with the disjunction "horses or cows and moose at dusk or cardboard horses or . . . " If so, why should we not say that when you are looking at the horses in the field, your belief has the *disjunctive* content "there are horses *or* cows *or* moose at dusk *or* cardboard horses *or* robot horses in the field"? This so-called disjunction problem has turned out to be a recalcitrant difficulty for the causal-correlational approach; it has been actively discussed, but there seems no solution that commands a consensus.[9]
4. We seem to have direct and immediate knowledge of what we believe, desire, and so on. I know, directly and without having to depend on evidence, that I believe it will rain tomorrow. That is, I seem to have direct knowledge of the content of my beliefs. There may be exceptions, but that does not overturn the general point. According to the correlational approach, my belief that there are horses in the field has the content it has because it correlates, or covaries, with the presence of horses in my vicinity. But this correlation is not some-

> thing that I know directly, without evidence or observation. So the correlational approach appears inconsistent with the special privileged status of our knowledge of the contents of our own mental states. (We discuss this issue further later, in connection with content externalism.)

These are some of the initial issues and difficulties for the correlational approach; whether, or to what extent, these difficulties can be overcome without compromising the naturalistic-reductive spirit of the theory remains an open question. Quite possibly, most of the difficulties are not really serious and can be resolved by further elaborations and supplementations. It may well be that this approach is the most promising one—in fact, the only viable one that promises to give a non-question-begging, naturalistic account of mental content.

Misrepresentation and the Teleological Approach

One important fact about representation is the possibility of *misrepresentation*. Misrepresentation does occur; you, or a mental-neural state of yours, may represent that there are horses in the field when there are none in sight. Or your perception may represent a red tomato in front of you when there is none (think about Macbeth and his bloody dagger). In such cases, misrepresentation occurs: The representational state misrepresents, and the representation is false. Representations have contents, and contents are "evaluable" in respect of truth, accuracy, fidelity, and related criteria of representational "success." It seems clear, then, that any account of representation must allow for the possibility of misrepresentation as well of course as correct, or successful, representation, just as any account of belief must allow for the possibility of false belief. One way of seeing how this could be a problem with the correlational approach is to go back to the disjunction problem discussed earlier. Suppose you form a representation with the content "there are horses over there" when there are no horses but only cows seen in the dusk. In such a case it would be natural to regard your purported representation as a misrepresentation—namely, as an instance of your representing something that does not exist, or representing something to be such and such when it is not such and such. But if we follow (C) literally, this seems impossible. If your representation was occasioned by cows seen in the dusk as well as horses, we would have to say that the representation has the content "horses or cows seen in the dusk" and that

that would make the representation correct and veridical. It would seem that (C) does not allow false beliefs or misrepresentations. But there surely are cases of misrepresentation; our cognitive systems are liable to produce false representations, even though they may be generally reliable.

This is where the teleological approach comes in to help out.[10] The basic concept employed in the teleological approach is that of a "function." For representation R to indicate (and thus represent) C, it is neither sufficient—nor necessary—that "whenever R occurs C occurs" holds. Rather, what must hold is that R has the *function* of indicating C—to put it more intuitively, R *is supposed* to indicate C and it is R's *job* to indicate C. Your representation has the content "there are horses over there" and not "there are horses or cows in the dusk over there" because it has the function of indicating the presence of horses, not horses or cows in the dusk. But things can go wrong, and systems do not always perform as they are supposed to. You form a representation of horses in the absence of horses; such a representation is *supposed* to be formed only when horses are present. That is exactly what makes it a case of misrepresentation. So it seems that the correlational-causal approach suitably supplemented with reference to function could solve the problem of misrepresentation.

But how does a state of a person or organism acquire a function of this kind? It is easy enough to understand function talk in connection with artifacts because we can invoke the purposes and intentions of their human designers and users. A thermometer reads 30°C, when the temperature is 20°C. What makes this a case of misrepresentation is that the thermometer's function is to indicate current air temperature, which is 20°C. That is the way the thermometer was designed to work and the way it is expected to work. It is the purposes and expectations external to the thermometer that give sense to the talk of functions. But this is something that we are not able to say, at least literally, about representations of natural systems, like humans and other higher animals. What gives a mental state (or a neural state) in us the function of representing some particular object or state of affairs? What gives a natural representation the job of representing "horses" rather than "horses or cows in the dusk"?

Philosophers who favor the teleological approach attempt to explain function in terms of evolution and natural selection. To say that representation R has the function of indicating C is to say that R has been selected, in the course of the evolution of the species to which the organism belongs, for the job of indicating C. This is like the fact that the heart has the function of pumping blood, or that the pineal gland has the function of secreting melatonin, because these organs have evolved for their performance of these tasks. Proper performance of these tasks presumably conferred adaptive advantages to our

ancestors. Similarly, we may presume that if R's function is to indicate C, performance of this job has given our ancestors biological advantages and, as some philosophers put it, R has been "recruited" by the evolutionary process to perform this function.

Exactly how the notion of function is to be explained is a further question that appears relatively independent of the core idea of the teleological approach. There are various and diverse biological-evolutionary accounts of function in the literature (see "For Further Reading" at the end of this chapter). Even if the theory of evolution were false and all biological organisms, including us, were created by God (so that we are God's "artifacts"), something like the teleological approach could still be right. It is God who gave our representations the indicating functions they have. But almost all contemporary philosophers of mind and of biology are naturalists, and it is important to them that function talk does not need to involve references to supernatural or transcendental plans, purposes, or designs. That is why they appeal to biology, learning and adaptation, and evolution for an account of function.

Narrow Content and Wide Content: Content Externalism

One thing that the correlational account of mental content highlights is this: Content has a lot to do with what is going on in the world, outside the physical boundaries of the creature. As far as what goes on inside is concerned, the frog in our world and the other-worldly frog are indistinguishable—they are in the same neural-sensory state, both registering a moving black dot. But in describing the representational content of their states, or what they "see," we advert to the conditions in the environments of the frogs: One frog sees a fly and the other sees a schmy. Or consider a simpler case: Peter is looking at a tomato, and Mary is also looking at one (a different tomato, but we suppose that it looks pretty much the same as Peter's tomato). Mary thinks to herself, "This tomato has gone bad," and Peter too thinks, "This tomato has gone bad." From the internal point of view, Mary's perceptual experience is indistinguishable from Peter's (we may suppose their neural states too are relevantly similar), and they would express their thoughts using the same words. But it is clear that the contents of their beliefs are different. For they involve different objects: Mary's belief is about the tomato she is looking at, and Peter's belief is about a different object altogether. Moreover, Mary's belief may be true and Peter's false, or vice versa. On one standard understanding of the notion of "content," beliefs with the same content must be true together or false together (that is,

contents serve as "truth conditions"). Obviously, the fact that Peter's and Mary's beliefs have different content is due to facts external to them; the difference in content cannot be explained in terms of what is going on inside the perceivers. It seems, then, that at least in this and other similar cases belief contents are differentiated, or "individuated," by reference to conditions external to the believer.

Beliefs whose content is individuated in this way are said to have "wide" or "broad" content. In contrast, beliefs whose content is individuated solely on the basis of what goes on inside the persons holding them are said to have "narrow" content. Alternatively, we may say that the content of an intentional state is narrow just in case it supervenes on the internal-intrinsic properties of the subject who is in that state, and that it is wide otherwise. This means that two individuals who are exactly alike in all intrinsic-internal respects must have the same narrow content beliefs but may well diverge in their wide content beliefs. Thus, our two frogs are exactly alike in internal-intrinsic respects but unlike in what their perceptual states represent. So the contents of these states do not supervene internally and are therefore wide.

Several well-known thought-experiments have been instrumental in persuading most philosophers that many, if not all, of our ordinary beliefs (and other intentional states) have wide content, that the beliefs and desires we hold are not simply a matter of what is going on inside our minds or heads. This is the doctrine of *content externalism*. Among these thought-experiments, the following two, the first due to Hilary Putnam and the second to Tyler Burge,[11] have been particularly influential.

Putnam's Thought-Experiment: Earth and Twin Earth

Imagine a planet, "Twin Earth," somewhere in the remote region of space, which is just like the Earth we inhabit, except in one respect: On Twin Earth, a certain chemical substance with the molecular structure XYZ, which has all the observable characteristics of water (it is transparent, dissolves salt and sugar, quenches thirst, puts out fire, freezes at 0°C, and so on), replaces water everywhere. So lakes and oceans on Twin Earth are filled with XYZ, not H_2O (that is, water), and Twin Earth people drink XYZ when they are thirsty, bathe and swim in XYZ, do their laundry in XYZ, and so on. Some Twin Earth people, including most of those who call themselves "Americans," speak English, which is indistinguishable from our English, and their use of the expression "water" is indistinguishable from its use on Earth.

But there is a difference: The Twin Earth "water" and our "water" refer to different things. When a Twin Earth inhabitant says, "Water is transparent," what she means is that XYZ is transparent. The same words when uttered by you, however, mean that water is transparent. The word "water" from a Twin Earth mouth means XYZ, not water, and the same word on your mouth means water, not XYZ. If you are the first visitor to Twin Earth and find out the truth about their "water," you may report back to your friends on Earth as follows: "At first I thought that the stuff that fills the oceans and lakes around here, and the stuff people drink and bathe in, was water, and it really looks and tastes just like water. But I just found out that it isn't water at all, although people around here call it 'water.' It's really XYZ, not water." You will not translate the Twin Earth word "water" into the English word "water"; you will translate it into "XYZ," or invent a new vernacular word, say "twater." We have to conclude then that the Twin Earth word "water" and our word "water" have different meanings, although what goes on inside the minds, or heads, of Twin Earth people may be exactly the same as what goes in ours, and their speech behavior involving their word "water" is indistinguishable from ours with our word "water." This semantic difference between our "water" and Twin Earth "water" is reflected in the way we describe and individuate mental states of people on Earth and people on Twin Earth. When a Twin Earth person says to the waiter, "Please bring me a glass of water!" she is expressing her desire for twater, and we will report, in *oratio obliqua*, that she wants some twater, not that she wants some water. When you say the same thing, you are expressing a desire for water, and we will say that you want water. You believe that water is wet, and your Twin Earth doppelganger believes that twater is wet. And so on. To summarize, people on Earth have water-thoughts and water-desires, whereas Twin Earth people have twater-thoughts and twater-desires; this difference is due to differences in the environmental factors external to the subjects, not to any differences in what goes on "inside" their heads.

Suppose we send an astronaut, Jones, to Twin Earth. She does not realize at first that the liquid she sees in the lakes and coming out of the tap is not water. She is offered a glass of this transparent liquid by her Twin Earth host and thinks to herself, "That's a nice, cool glass of water—just what I needed." Consider Jones's belief that the glass contains cold water. This belief is false, since the glass contains not water but XYZ, that is, twater. Although she is now on Twin Earth, in an environment full of twater and devoid of water, she is still subject to the standards current on Earth: Her words mean, and her thoughts are individuated, in accordance with the criteria that prevail on Earth. What

this shows is that a person's *past associations* with her environment play a role in determining her present meanings and thought contents. If Jones stays on Twin Earth long enough—say, a dozen years—we will likely interpret her word "water" to mean twater, not water, and attribute to her twater-thoughts rather than water-thoughts—that is, eventually she will come under the linguistic conventions of Twin Earth.

If these considerations are by and large correct, they show that two supervenience theses fail: First, the meanings of our expressions do not in general supervene on our *internal*, or *intrinsic*, physical-psychological states. I and my molecule-for-molecule-identical Twin Earth doppelganger are indistinguishable as far as our internal lives, both physical and mental, are concerned, and yet our words have different meanings—my "water" means water and his "water" means XYZ, that is, twater. Second, and this is what is of immediate interest to us, the contents of beliefs and other intentional states also fail to supervene on internal physical-psychological states. You have water-thoughts and your doppelganger has twater-thoughts, in spite of the fact that you two are in the same internal states, physical and psychological. Beliefs, or thoughts, are individuated by content—that is, that we regard beliefs with the same content as the same belief, and beliefs with different content count as different. So your water-thoughts and your twin's twater-thoughts are different thoughts. What beliefs you hold depends on your relationship, both past and present, to the things and events in your surroundings, as well as on what goes on inside you. The same goes for other content-carrying intentional states. If this is right, intentional states have wide content.

Burge's Thought-Experiment: Arthritis and "Tharthritis"

Consider a person, call him Peter, in two situations. (1) *The actual situation*: Peter thinks "arthritis" means inflammation of the bones. (It actually means inflammation of the bone joints.) Feeling pain and swelling in his thigh, Peter complains to his doctor, "I have arthritis in my thigh." His doctor tells him that people can have arthritis only in their joints. Two points should be noted: First, Peter believed, before he talked to his doctor, that he had arthritis in his thigh; and second, this belief was false.

(2) *A counterfactual situation*: Nothing has changed with Peter. Experiencing swelling and pain in his thigh, he complains to his doctor, "I have arthritis in my thigh." What is different about the counterfactual situation concerns the use of the word "arthritis" in Peter's speech community: In the situation we are imagining, the word is used to refer to inflammation of bones, not just bone

joints. That is, in the counterfactual situation Peter has a correct understanding of the word "arthritis," unlike in the actual situation. In the counterfactual situation, then, Peter is expressing a true belief when he says "I have arthritis in my thigh." But how should we report Peter's belief concerning the condition of his thigh in the counterfactual situation—that is, report in *our* language (in the actual world)? We cannot say that Peter believes that he has arthritis in his thigh, because in our language "arthritis" means inflammation of joints and he clearly does not have that, making his counterfactual belief false. We might coin a new expression (to be part of our language), "tharthritis," to mean inflammation of bones as well as of joints, and say that Peter, in the counterfactual situation, believes that he has tharthritis in his thigh. Again, note two points: First, in the counterfactual situation, Peter believes not that he has arthritis in his thigh but that he has tharthritis in his thigh; and second, this belief is true.

What this thought-experiment shows is that the content of belief depends, in part but crucially, on the speech practices of the linguistic community in which we situate the subject. Peter in the actual situation and Peter in the counterfactual situation are exactly alike when taken as an individual person (that is, when we consider his internal-intrinsic properties alone), including his speech habits (he speaks the same idiolect in both situations) and inner mental life. Yet he has different beliefs in the two situations: Peter in the actual world has the belief that he has arthritis in his thigh, which is false, but in the counterfactual situation he has the belief that he has tharthritis in his thigh, which is true. The only difference in the two situations is that of the linguistic practices of Peter's community (concerning the use of the word "arthritis"), not anything intrinsic to Peter himself. If this is right, beliefs and other intentional states do not supervene on the internal physical-psychological states of persons; if supervenience is wanted, we must include in the supervenience base the linguistic practices of the community to which people belong.

Burge argues, persuasively for most philosophers, that the example can be generalized to show that almost all contents are wide—that is, externally individuated. Take the word "brisket" (another of his examples): Some of us mistakenly think that brisket comes only from beef, and it is easy to see how a case analogous to the arthritis example can be set up. (The reader is invited to try.) As Burge points out, the same situation seems to arise for any word whose meaning is incompletely, or defectively, understood—in fact, any word whose meaning *could* be incompletely understood, which includes pretty much every word. When we profess our beliefs using such words, our beliefs are identified and individuated by the socially determined meanings of these

words (recall Peter and his "arthritis" in the actual situation), and a Burge-style counterfactual situation can be set up for each such word. Moreover, we seem to identify our own beliefs in terms of the words we would use to express them, even if we are aware that our understanding of these words is incomplete or defective. (How many of us know the correct meaning of, say, "mortgage," "justice of the peace," or "galaxy"?) This shows, it has been argued, that almost all of our ordinary belief attributions involve wide content.

If this is right, the question naturally arises: Are there beliefs whose content is not determined by external factors? That is, are there beliefs with "narrow content"? There appear to be beliefs, and other intentional states, that do not imply the existence of anything, or do not refer to anything, outside the subject who has them. For example, Peter's belief that he is in pain or that he exists, or that there are no unicorns, does not require anything other than Peter to exist, and it would seem that the content of these beliefs is independent of conditions external to Peter. If so, the narrowness of these beliefs is not threatened by considerations of the sort that emerged from the Twin Earth thought-experiment. But what of Burge's arthritis thought-experiment? Consider Peter's belief that he is in pain. Could we run on the word "pain" Burge's argument on "arthritis"? Surely it is possible for someone to misunderstand the word "pain" or any other sensation term. Suppose Peter thinks that "pain" applies to both pains and severe itches and that on experiencing a bad itch on his shoulder, he complains to his wife about an annoying "pain" in the shoulder. If the Burge-style considerations apply here, we have to say that Peter is expressing his belief that he is having a pain in his shoulder and that this is a false belief.

The question is whether that is indeed what we would, or should, say. It would seem not unreasonable that knowing what we know about Peter's misunderstanding of the word "pain" and the sensation he is actually experiencing, the correct thing to say is that he believes, and in fact knows, that he is experiencing an itch on his shoulder. It is only that in saying, "I am having a pain in my shoulder," he is misdescribing his sensation and hence misreporting his belief.

Now, consider the following counterfactual situation: In the linguistic community to which Peter belongs, "pain" is used to refer to pains and severe itches. How would we report, in our own words, the content of Peter's belief in the counterfactual situation when he utters "I have a pain in my shoulder"? Remember that both in the actual and counterfactual situations, Peter is having a bad itch, and no pain. There are these possibilities: (i) We say "He believes that he has a pain in his shoulder"; (ii) we say "He believes that he has a

bad itch in his shoulder"; and (iii) we do not have a word in English that can be used for expressing the content of his belief (but we could introduce a neologism, "painitch," and say "Peter believes that he is having a painitch in his shoulder"). Obviously, (i) has to be ruled out; if (iii) is what we should say, the arthritis argument applies to the present case as well, since this would show that a change in the social environment of the subject can change the belief content attributed to him. But it is not obvious that this, rather than (ii), is the correct option. It seems to be an open question, then, whether the arthritis argument applies to cases involving beliefs about one's own sensations, and there seems to be a reason for the inclination to say of Peter in the actual world that he believes he is having severe itches rather than that he believes he is having pains. The reason is that if we were to opt for the latter, it would make his belief false, and this is a belief about his own current sensations. But we assume that under normal circumstances people do not make mistakes in identifying their current sensory experiences. This assumption need not be taken as a contentious philosophical doctrine; arguably, recognition of first-person authority on such matters also reflects our common social-linguistic practices, and this may very well override the kinds of considerations advanced by Burge in the case of arthritis and the rest.

These considerations should give us second thoughts concerning Burge's thought-experiment involving arthritis and tharthritis. As you will recall, this involved a person, Peter, who misunderstands the meaning of "arthritis" and, on experiencing pain in his thigh, says to his doctor, "I have arthritis in my thigh." With Burge, we said that Peter believes that he has arthritis in his thigh, and that this belief is false. Is this what we should really say? Isn't there an option, perhaps a more reasonable one, of saying that Peter, in spite of the words he used, doesn't believe that he has arthritis in his thigh; rather, the content of the belief he expresses when he says to the doctor "I have arthritis in my thigh" is to the effect that he has pain in his thigh, or that he has an inflammation of his thigh bone. He does have a false, or defective, belief—about the meaning of the word "arthritis"—and this leads him to misreport the content of his belief. Of course, it is no surprise that the meanings of words depend on the linguistic practice of the speech community. The reader is invited to ponder this way of responding to Burge's thought-experiment.

Another point to consider is beliefs of animals without speech. Do cats and dogs have beliefs and other intentional states whose contents can be reported in the form: "Fido believes that *p*," where *p* stands in for a declarative sentence? We do say things like "Fido believes that Charlie is calling him to come upstairs," "He believes that the mail carrier is at the door," and so on. But it is clear

that the arthritis-style arguments cannot be applied to such beliefs since Fido does not belong to any speech community and the only language that is involved is our own, namely, the language of the person who makes such belief attributions. In what sense, then, could animal beliefs be externally individuated? It seems that Putnam's Twin Earth–style considerations can be applied to animal beliefs (also recall our fly-schmy example), but Burge-style argument cannot. However, the case of animal beliefs can cut both ways as far as Burge's argument is concerned, for we might argue, as some philosophers have,[12] that nonlinguistic animals are not capable of having intentional states (in particular, beliefs) and, therefore, the inapplicability of Burge's considerations is only to be expected. Some will find this line of thinking highly implausible, namely that only animals that use language for social communication are capable of having beliefs and other intentional states.

The Metaphysics of Wide Content States

Considerations involved in the two thought-experiments show that many, if not all, of our ordinary beliefs and other intentional states have wide content. Their contents are "external": They are determined, in part but importantly, by factors outside the subject—factors in her physical and social environment and in her history of interaction with it. Before these externalist considerations were brought to our attention, philosophers used to think that beliefs, desires, and the like were "in the mind," or at least "in the head." Putnam, the inventor of the Twin Earth parable, declared, "Cut the pie any way you like, 'meanings' just ain't in the head."[13] Should we believe that beliefs and desires are not in the head, or in the mind, either? If so, where are they? *Outside* the head? If so, just where? Does that even make sense? Let us consider some possibilities.

1. We might say that the belief that water and oil do not mix is constituted in part by water and oil—that the belief itself, in some sense, involves the actual stuff, water and oil, in addition to the person (or her "head") having the belief. A similar response in the case of arthritis would be that Peter's belief that he has arthritis is in part constituted by his linguistic community. The general idea is that all the factors that play a role in determining the content of a belief *ontologically constitute* that belief; the belief is a state that comprises these items within itself. Thus, we have a simple explanation for just how your belief that water is wet differs from your Twin Earth doppelganger's belief that twater is wet: Yours includes water as a con-

stituent, and hers includes twater as a constituent. On this approach, then, beliefs extrude from the subject's head into the world, and there are no bounds to how far they can reach. The whole universe would, on this approach, be a constituent of your beliefs about the universe! Moreover, all beliefs about the universe would appear to have exactly the same constituent, namely, the universe. This sounds absurd, and it is absurd. We can also see that this general approach would make the causal role of beliefs difficult to understand—beliefs as either causes or effects.

2. We might consider the belief that water and oil do not mix as a *relation* holding between the subject, on the one hand, and water and oil, on the other. Or alternatively, we take the belief as a *relational property* of the subject involving water and oil. (That Socrates is married to Xanthippe is a relational fact; Socrates also has the relational property of being married to Xanthippe, and conversely, Xanthippe has the relational property of being married to Socrates.) This approach makes causation of beliefs more tractable: We can ask, and will sometimes be able to answer, how a subject came to bear this belief relation to water and oil, just as we can ask how Xanthippe came to have the relational property of being married to Socrates. But what of other determinants of content? As we saw, belief content is determined in part by the history of one's interaction with one's environment. And what of the social-linguistic determinants, as in Burge's examples? It seems at least awkward to consider beliefs as relations with respect to these factors.
3. The third possibility is to consider beliefs to be wholly internal to the subjects who have them but consider their contents, when they are wide, as giving *relational specifications*, or *descriptions*, of the contents. On this view, beliefs may be neural states or other types of physical states of organisms to which they are attributed, and as such they are "in" the believer's head, or mind. Contents, then, are construed as ways of specifying, or describing, the representational properties of these states; wide contents are thus specifications in terms that involve factors and conditions external to the subject, both physical and social, both current and historical. We can refer to, or pick out, Socrates by relational descriptions, that is, in terms of his relational properties—for example, "the husband of Xanthippe," "the Greek philosopher who drank hemlock in a prison in Athens," "Plato's mentor," and so on. But this does not mean that

> Xanthippe, hemlock, or Plato is a constituent part of Socrates, nor does it mean that Socrates is some kind of a "relational entity." Similarly, when we specify Jones's belief as the belief that water and oil do not mix, we are specifying this belief relationally, by reference to water and oil, but this does not mean that water and oil are constituents of the belief or that the belief itself is a relation to water and oil.

Let us look at this last approach in a bit more detail. Consider physical magnitudes such as mass and length, which are standardly considered to be paradigm examples of intrinsic properties of material objects. How do we *specify*, *represent*, or *measure* the mass or length of an object? The answer: relationally. To say that this metal rod has a mass of three kilograms is to say that it bears a certain relationship to the International Prototype Kilogram. (It would balance, on an equal-arm balance, three objects that each balance the Standard Kilogram.) Likewise, to say that the rod has a length of two meters is to say that it is twice the length of the Standard Meter (or twice the distance traveled by light in a vacuum in a certain specified fraction of a second). These properties, mass and length, are intrinsic, but their specifications or representations are extrinsic and relational, involving relationships to other things and properties in the world. Moreover, the availability of such extrinsic representations may be essential to the utility of these properties in the formulation of scientific laws and explanations. They make it possible to relate a given intrinsic property to other significant properties in theoretically interesting and fruitful ways. Similar considerations might explain the usefulness of wide contents, or relational descriptions of beliefs, in vernacular explanations of human behavior.

In physical measurements, we use numbers to specify properties of objects, and these numbers involve relationships to other objects (see the above discussion of what "three kilograms" refers to). In attributing to persons beliefs, we use propositions, or content sentences, to specify their contents, and these propositions often involve references to objects and events outside the believers. When we say that Jones believes that water is wet, we are using the content sentence "water is wet" to specify this belief, and the appropriateness of this sentence as a specification of the belief depends on Jones's relationship, past and present, to her environment. What Burge's examples show is that the choice of a content sentence may depend also on the social-linguistic facts about the person holding the belief. In a sense, we are "measuring" people's mental states using sentences, just as we measure physical magnitudes using numbers.[14] Just as the assignment of numbers in measurement involves relationships to things

other than the things whose magnitudes are being measured, the use of content sentences in the specification of belief contents makes use of, and depends on, factors outside the subject. In both cases the informativeness and utility of the specifications—the assigned numbers or sentences—depend crucially on the involvement of external factors and conditions.[15]

This approach seems to have much to recommend itself over the other two. It locates beliefs and other intentional states squarely within the subjects; ontologically, they are states of the persons holding them, not something that somehow extrudes from them into the outside, like some green goo we see in science fiction films! This is a more elegant metaphysical picture than its alternatives. What is "wide" about these states is their specifications or descriptions, not the states themselves. And there are good reasons for using wide content specifications. For one, we want them to indicate the representational contents of beliefs (and other intentional states)—what states of affairs they represent—and it is no surprise that this involves reference to external conditions. After all, the whole point of beliefs is to represent states of affairs in the world, outside the believer. For another, the sorts of social-linguistic constraints involved in Burge's examples may be crucial to the uniformity, stability, and intersubjectivity of content attributions. The upshot is that it is important not to conflate the ontological status of intentional states with the modes of their specification.

Is Narrow Content Possible?

You believe that water extinguishes fires, and your twin on Twin Earth believes that twater extinguishes fires. The two beliefs have different contents: What you believe is not the same as what your twin believes. But leaving the matter here is unsatisfying; it misses something important—something *psychologically* important—that you and your twin share in holding these beliefs. "Narrow content" is supposed to capture this something you and your twin share.

First, we seem to have a strong sense that both you and your twin conceptualize the same state of affairs in holding the beliefs about water and twater, respectively; the way things seem to you when you think that freshwater fills the Great Lakes must be the same, we feel, as the way things seem to your twin when she thinks that fresh twater fills the Twin Earth Great Lakes. From an internal psychological perspective, your thought and her thought seem to have the same significance. In thinking of water, you perhaps have the idea of a substance that is transparent, flows a certain way, tastes a certain way, and so on; in thinking of twater, your twin has the same associations. Or take the frog case: Isn't it plausible to suppose that the frog in our world that detects a fly and the

other-worldly frog that detects a schmy are in the same perceptual state—a state whose "immediate" content consists in a black dot flitting across the visual field? There is a strong intuitive pull toward the view that there is something important that is common to your psychological life and your twin's, and to our frog's perceptual state and the other-worldly frog's, that could reasonably be called "content."

Second, consider your behavior and your twin's behavior: They show a lot in common. For example, when you find your couch on fire, you pour water on it; when your twin finds her couch on fire, she pours twater on it. If you were visiting Twin Earth and found a couch on fire there, you would pour twater on it too (and conversely, if your twin is visiting Earth). In ordinary situations your behavior involving water is the same as her behavior involving twater; moreover, your behavior would remain the same if twater were substituted for water everywhere, and this goes for your twin as well mutatis mutandis. It seems then that the water-twater difference is *psychologically irrelevant*—irrelevant for behavior causation and explanation. The difference between water-thoughts and twater-thoughts cancels itself out, so to speak. What is important for psychological explanation seems to be what you and your twin share, namely, thoughts with narrow content. So the question arises: Does psychological theory need wide content? Can it get by with narrow content alone?

We have seen some examples of beliefs that plausibly do not depend on the existence of anything outside the subject holding them: your beliefs that you exist, that you are in pain, that unicorns do not exist, and the like. Although we have left open the question of whether the arthritis argument applies to them, they are at least "internal" or "intrinsic" to the subject in the sense that for these beliefs to exist, nothing outside the subject needs to exist. It appears, then, that these beliefs do not involve anything external to the believer and therefore that these beliefs supervene solely on the factors internal to him (again barring the possibility that the Burge-style considerations generalize to all expressions without exception).

However, a closer look reveals that some of these beliefs are not supervenient only on internal states of the believer. For we need to consider the involvement of the subject herself in the belief. Consider Mary's belief that she is in pain. The content of this belief is that she—that is, Mary—is in pain. This is the state of affairs represented by the belief, and this belief is true just in case that state of affairs obtains—that is, just in case Mary is in pain. Now we put Mary's twin on Twin Earth in the same internal physical state that Mary is

in when she has this belief. If mind-body supervenience, as *intuitively understood*, holds, it would seem that Mary's twin too will have the belief that she is in pain. However, her belief has the content that *she* (Twin Earth Mary) is in pain, not that Mary is in pain. The belief is true if and only if Mary's twin is in pain. Beliefs with the same content are true together, or false together. It follows, then, that belief contents in cases of this kind do not supervene on the internal-intrinsic physical properties of persons. This means that the following two ideas that are normally taken to lie at the core of the notion of "narrow content" fail to coincide: (1) Narrow content is internal and intrinsic to the believer and does not involve anything outside her current state; and (2) narrow content, unlike wide content, supervenes on the current internal physical state of the believer.[16]

One possible way to look at the situation is this: What examples of this kind show is not that these beliefs do not supervene on the internal physical states of the believer, but rather that we should revise the notion of "same belief"—that is, we need to revise the criteria of belief individuation. In our discussion thus far, individual beliefs (or "belief tokens") have been considered to be "the same belief" (or the same "belief type") just in case they have the same content; on this view, two beliefs have the same content only if their truth condition is the same (that is, necessarily they are true together or false together). As we saw, Mary's belief that she, Mary, is in pain and her twin's belief that she, the twin Mary, is in pain do not have the same truth condition and hence must count as belonging to different belief types. That is why supervenience fails for these beliefs. However, there is an obvious and natural sense in which Mary and her twin have "the same belief"—even beliefs with "the same content"—when each believes that she is in pain. More work, however, needs to be done to capture this notion of content or sameness of belief,[17] and that is part of the project of explicating the notion of narrow content.

As noted, it is widely accepted that most of our ordinary belief attributions, as well as attributions of other intentional states, involve wide content. Some hold not only that all contents are wide but that the very notion of narrow content makes no sense. One point often made against narrow content is its alleged ineffability: How do we capture the shared content of Jones's belief that water is wet and her twin's belief that twater is wet? And if there is something shared, why is it a kind of "content"?

One way the friends of narrow content have tried to deal with such questions is to treat narrow content as an abstract technical notion, roughly in the following sense. The thing that Mary and her twin share plays the following

role: If anyone has it and has acquired her language on Earth (or in an environment containing water), her word "water" refers to water and she has water-thoughts; if anyone has it and has acquired her language on Twin Earth (or in an environment containing twater), her word "water" refers to twater and she has twater-thoughts; for anyone who has it and has acquired her language in an environment in which a substance with molecular structure PQR replaces water everywhere, her word "water" refers to PQR; and so on. The same idea applies to the frog case: What the two frogs, one in this world and the other in a world with schmies but no flies, have in common is this: If a frog has it and inhabits an environment with flies, it has the capacity to have flies as part of its perceptual content, and similarly for frogs in a schmy-inclusive environment. Technically, narrow content is a function from environmental contexts (including contexts of language acquisition) to wide contents (or truth conditions).[18] One question that has to be answered is why narrow content in that sense is a kind of content. For isn't it true, by definition, that content is "semantically evaluable"—that is, that it is something that can be true or false, accurate to various degrees, and so on? Narrow content, conceived as a function from environment to wide content, does not seem to meet this conception of content; it does not seem like the sort of thing that can be said to be true or false. Here various strategies for meeting this point seem possible; however, whether any of them will work is an open question.

Two Problems for Content Externalism

We briefly survey here two outstanding issues confronting the thesis that most, perhaps all, of our intentional mental states have wide content. (The first was briefly alluded to earlier.)

The Causal-Explanatory Efficacy of Wide Content

Even if we acknowledge that commonsense psychology individuates intentional states widely and formulates causal explanations of behavior in terms of wide content states, we might well ask whether this is an ineliminable feature of such explanations. Several considerations can be advanced to cast doubt on the causal-explanatory efficacy of wide content states. First, we have already noted the similarity between the behaviors of people on Earth and those of their Twin Earth counterparts in relation to water and twater, respectively. We saw that in formulating causal explanations of behaviors, the difference between water-thoughts and twater-thoughts somehow cancels itself out

by failing to manifest itself in a difference in the generation of behavior. Second, to put the point another way, if you are a psychologist who has already developed a working psychological theory of people on Earth, formulated in terms of content-bearing intentional states, you obviously would not start all over again from scratch when you want to develop a psychological theory for Twin Earth people. In fact, you are likely to say that people on Earth and those on Twin Earth have "the same psychology"—that is, the same psychological theory holds for both groups. In view of this, isn't it more appropriate to take the difference between water-thoughts and twater-thoughts, or water-desires and twater-desires, merely as a difference in the values of a contextual parameter to be fixed to suit the situations to which the theory is applied rather than as an integral element of the theory itself? If this is correct, doesn't wide content drop out as part of the theoretical apparatus of psychological theory?

Moreover, there is a metaphysical point to consider: The proximate cause of my physical behavior (say, my bodily motions), we feel, must be "local"—it must be a series of neural events originating in my central nervous system that causes the contraction of appropriate muscles, which in turn moves my limbs. This means that what these neural events represent in the outside world is irrelevant to behavior causation: If the same neural events occur in a different environment so that they have different representational (wide) content, they would still cause the same physical behavior. That is, we have reason to think that proximate causes of behavior are *locally* supervenient on the internal physical states of an organism, but that wide content states are not so supervenient. Hence, the wideness of wide content states is not relevant to causal explanations of physical behavior. (You may recall discussion of the irrelevance of representational contents of computational states to the course of computational process, in chapter 5.)

One way in which the friends of wide content have tried to counter these considerations goes as follows. What we typically attempt to explain in commonsense psychology is not physical behavior but action—not why your right hand moved thus and so, but why you turned on the stove, why you boiled the water, why you made the tea. To explain why your hand moved in a certain way, it may suffice to advert to causes "in the head," but to explain why you turned on the stove or why you boiled the water, we must invoke wide content states: because you wanted to heat the kettle of water, because you wanted to make a cup of tea for your friend, and so on. Behaviors explained in typical commonsense explanations are given under "wide descriptions," and we need wide content states to explain them. So the point of the reply is that we need wide content to explain "wide behavior." Whether this response is

sufficient is something to think about. In particular, we might raise questions as to whether the wideness of thoughts and the wideness of behavior are playing any real role in the causal-explanatory relation involved, or whether they merely ride piggyback, so to speak, on an underlying causal-explanatory relationship between the neural states, or narrow content states, and physical behavior. (The issues discussed in an earlier section, "The Metaphysics of Wide Content States," are directly relevant to these causal-explanatory questions about wide content. The reader is encouraged to think about whether the third option described in that section could help the content externalist to formulate a better response.)

Wide Content and Self-Knowledge

How do we know that Mary believes that water is wet and that Mary's twin on Twin Earth believes that twater is wet? Because we know that Mary's environment contains water and that Mary's twin's environment contains twater. Now consider the matter from Mary's point of view: How does she know that she believes that water is wet? How does she know the content of her own thoughts?

We believe that a subject has special, direct access to her own mental states (see chapters 1 and 9). Perhaps the access is not infallible and does not extend to all mental states, but it is uncontroversial that there is special first-person authority in regard to one's own occurrent thoughts. When you reflect on what you are thinking, you apparently know directly, without further evidence or reasoning, what you think; the content of your thought is immediately and directly accessible to you, and the question of having evidence or doing research does not arise. If you think that the shuttle bus is late and you might miss your flight, you know, in the very act of thinking, that that is what you are thinking. First-person knowledge of the contents of one's own current thoughts is direct and immediate and carries a special sort of authority.

Return now to Mary and her knowledge of the content of her belief that water is wet. It seems plausible to think that in order for her to know that her thought is about water, not about twater, she is in the same epistemic situation that we are in with respect to the content of her thought. We know that her thought is about water, not twater, because we know, from observation, that her environment is water-inclusive, not twater-inclusive. But why doesn't she too have to know that if she is to know that her thought is about water, not twater, and how can she know something like that without observation or evidence? It looks like she may very well lose her specially privileged epistemic ac-

cess to the content of her own thought, because her knowledge of her thought content is now put on the same footing as third-person knowledge of it.

To make this more vivid, suppose that Twin Earth exists in a nearby planetary system and we can travel between Earth and Twin Earth. It is plausible to suppose that if one spends a sufficient amount of time on Earth (or Twin Earth), one's word "water" becomes locally acclimatized and begins to refer to the local stuff, water or twater, as the case may be. Now, Mary, an inveterate space traveler, forgets on which planet she has been living for the past several years, whether it is Earth or Twin Earth; surely that is something she cannot know directly without evidence or observation. Now ask: Can she know, directly and without further investigation, whether her thoughts (say, the thought she expresses when she mutters to herself, "The tap water in this fancy hotel doesn't taste so good") are about water or twater? It prima facie makes sense to think that just as she cannot know, without additional evidence, whether her present use of the word "water" refers to water or twater, she cannot know, without investigating her environment, whether her thought, on seeing the steaming kettle, has the content that the water is boiling or that the twater is boiling. If something like this is right, then content externalism would seem to have the consequence that most of our knowledge of our own intentional states is not direct and, like most other kinds of knowledge, must be based on evidence. That is to say, content externalism appears to be prima facie incompatible with privileged first-person access to one's own mind. Content externalists are, of course, not without answers, but an examination of these is beyond the scope of this chapter.

* * *

These issues concerning wide and narrow content—especially the second concerning content externalism and self-knowledge—have been vigorously debated and are likely to be with us for some time. Their importance can hardly be exaggerated: Content-carrying states—that is, intentional states like belief, desire, and the rest—constitute the central core of our commonsense ("folk") psychological practices, providing us with a framework for formulating explanations and predictions of what we and our fellow humans do. Without this essential tool for understanding and anticipating human action and behavior, a communal life would be unthinkable. Moreover, the issues go beyond commonsense psychology. There is, for example, this important question about scientific psychology and cognitive science: Should the sciences of human behavior and cognition make use of content-carrying intentional

states like belief and desire, or their more refined and precise scientific analogues, in formulating its laws and explanations? Or should they, or could they, transcend the intentional idiom by couching their theories and explanations in purely nonintentional (perhaps, ultimately neurobiological) terms? These questions concern the centrality of content-bearing, representational states to the explanation of human action and behavior—both in everyday psychological practices and in theory construction in scientific psychology.

For Further Reading

On interpretation theory, see the works by Davidson, Quine, and Lewis cited in footnote 1; see also Daniel C. Dennett, "Intentional Systems" and "True Believers."

On causal-correlational theories of content, see the works cited in footnote 7; see also Robert Cummins, *Meaning and Mental Representation*, especially chapters 4 through 6. Another useful book on issues of mental content, including some not discussed in this chapter, is Lynne Rudder Baker, *Explaining Attitudes*. There are several helpful essays in *Meaning in Mind*, edited by Barry Loewer and Georges Rey.

On teleological accounts of mental content, see Fred Dretske, "Misrepresentation," and Ruth Millikan, "Biosemantics." Karen Neander's "Teleological Theories of Mental Content" is a comprehensive survey and analysis.

On narrow and wide content, the two classic texts that introduced the issues are Hilary Putnam, "The Meaning of 'Meaning,'" and Tyler Burge, "Individualism and the Mental." See also Fodor's *Psychosemantics* and "A Modal Argument for Narrow Content." On narrow content, see Gabriel Segal, *A Slim Book About Narrow Content*. For a discussion of these issues in relation to scientific psychology, see Frances Egan, "Must Psychology Be Individualistic?" Joseph Mendola's *Anti-Externalism* is an extended and helpful analysis and critique of externalism; see chapter 2 for discussion of Putnam's and Burge's thought-experiments in support of externalism.

Concerning content and causation, the reader may wish to consult the following: Colin Allen, "It Isn't What You Think: A New Idea About Intentional Causation"; Lynne Rudder Baker, *Explaining Attitudes*; Tim Crane, "The Causal Efficacy of Content: A Functionalist Theory"; Fred Dretske, *Explaining Behavior* and "Minds, Machines, and Money: What Really Explains Behavior"; Jerry Fodor, *Psychosemantics* and "Making Mind Matter More"; and Pierre Jacob, *What Minds Can Do*.

On wide content and self-knowledge, see Donald Davidson, "Knowing One's Own Mind"; Tyler Burge, "Individualism and Self-Knowledge"; Paul Boghossian, "Content and Self-Knowledge"; and John Heil, *The Nature of True Minds*, chapter 5. Three recent collections of essays on the issue are *Externalism and Self-Knowledge*, edited by Peter Ludlow and Norah Martin; *Knowing Our Own Minds*, edited by Crispin Wright, Barry C. Smith, and Cynthia Macdonald; and *New Essays on Semantic Externalism and Self-Knowledge*, edited by Susan Nuccetelli.

Notes

1. The discussion in this section is based on the works of W. V. Quine and Donald Davidson—especially Davidson's. See Quine on "radical translation" in his *Word and Object*, chapter 2. Davidson's principal essays on interpretation are included in his *Inquiries into Truth and Interpretation*; see, in particular, "Radical Interpretation," "Thought and Talk," and "Belief and the Basis of Meaning." Also see David Lewis, "Radical Translation."

2. Here we are making the plausible assumption that we can determine, on the basis of observation of Karl's behavior, that he affirmatively utters, or holds true, a sentence S, without our knowing what S means or what belief Karl expresses by uttering S. (The account would be circular otherwise.) It can be granted that holding true a sentence is a psychological attitude or event. For further discussion of this point, see Davidson, "Thought and Talk," pp. 161–162.

3. The parenthetical part is often assumed without being explicitly stated. Some writers state it as a separate principle, sometimes called the "requirement of rationality." There are many inequivalent versions of the charity principle in the literature. Some restrictions on the class of beliefs to which charity is to be bestowed are almost certainly necessary. For our examples, all we need is to say that speakers' beliefs about observable features of their immediate environment are generally true; that is, we restrict the application of charity to "occasion sentences" whose utterances are sensitive to the observable change in the environment.

4. Such a position seems implicit in, for example, Daniel Dennett's "True Believers."

5. The following statement from Davidson, who has often avowed himself to be a mental realist, seems to have seemingly irrealist, or possibly relativist, implications: "For until the triangle is completed connecting two creatures [the interpreter and the subject being interpreted], and each creature with

common features of the world, there can be no answer to the question whether a creature, in discriminating between stimuli, is discriminating stimuli at sensory surfaces or somewhere further out, or further in. Without this sharing of reactions to common stimuli, thought and speech would have no particular content—that is, no content at all. It takes two points of view to give a location to the cause of a thought, and thus, to define its content." See Davidson, "Three Varieties of Knowledge," pp. 212–213.

6. To use Robert Stalnaker's term in his *Inquiry*, p. 18. Fred Dretske, too, uses "indicator" and its cognates for similar purposes in his writings on representation and content.

7. This version captures the gist of the correlational approach, which has many diverse versions. Important sources include Fred Dretske, *Knowledge and the Flow of Information* and "Misrepresentation"; Robert Stalnaker, *Inquiry*; and Jerry A. Fodor, *Psychosemantics* and *A Theory of Content and Other Essays*. Dennis Stampe is usually credited with initiating this approach in "Toward a Causal Theory of Linguistic Representation." For discussion and criticisms, see Brian McLaughlin, "What Is Wrong with Correlational Psychosemantics?" (to which I am indebted in this section); and Louise Antony and Joseph Levine, "The Nomic and the Robust"; Lynne Rudder Baker, "Has Content Been Naturalized?"; and Paul Boghossian, "Naturalizing Content" in *Meaning in Mind*, ed. Barry Loewer and Georges Rey.

8. This means that S is entertaining this belief, actively in some sense, at the time.

9. For discussion of this issue, see the works cited in note 7.

10. This is not to say that the teleological approach is necessarily the only solution to the problem of misrepresentation or the disjunction problem. See Jerry A. Fodor, *A Theory of Content and Other Essays*.

11. Hilary Putnam, "The Meaning of 'Meaning'"; Tyler Burge, "Individualism and the Mental." The terms "narrow" and "wide" are due to Putnam.

12. Most notably Descartes and Davidson. See Davidson's "Rational Animals."

13. Hilary Putnam, "The Meaning of 'Meaning,'" p. 227.

14. This idea was first introduced by Paul M. Churchland in "Eliminative Materialism and the Propositional Attitudes." It has been systematically elaborated by Robert Matthews in "The Measure of Mind." However, these authors do not relate this approach to the issues of content externalism. For another perspective on the issues, see Ernest Sosa, "Between Internalism and Externalism."

15. Burge makes this point concerning content sentences in "Individualism and the Mental."

16. Beliefs with wide content will generally not supervene on the internal, intrinsic physical properties of the subjects. That is not surprising; the present case is worth noting because it apparently involves narrow content.

17. In this connection, see Roderick Chisholm's theory in *The First Person*, which does not take beliefs as relations to propositions but construes them as attributions of properties. David Lewis has independently proposed a similar approach in "Attitudes *De Dicto* and *De Se*." On an approach of this kind, both Mary and twin Mary are self-attributing the property of being in pain, and the commonality shared by the two beliefs consists in the self-attribution of the same property, namely that of being in pain.

18. See Stephen White, "Partial Character and the Language of Thought," and Jerry A. Fodor, *Psychosemantics*. See also Gabriel Segal, *A Slim Book About Narrow Content*.

CHAPTER 9

What Is Consciousness?

Nothing could be more familiar to us than the phenomenon of consciousness. We are conscious at every moment of our waking lives; it is a ubiquitous and unsurprising feature of everyday existence—except when we are in deep sleep, in a coma, or otherwise, well, unconscious. In one of its senses, "conscious" is just another word for "awake" or "aware," and we know what it is to be awake and aware—to awaken from sleep, general anesthesia, or a temporary loss of consciousness caused by a trauma to the head, and regain an awareness of what is going on in and around us.

Consciousness is a central feature of mentality—or at any rate the kind of mentality that we possess and value. A brain-dead person has suffered an irreversible loss of consciousness, and that seems the primary reason why brain death matters to us, personally and ethically. Most of us would be inclined to believe that for all human intents and purposes, a person who has permanently lost the capacity for consciousness is no longer with us. This suggests that consciousness might be a precondition of mentality and personhood—that any creature with mentality must be a conscious being. By any measure, consciousness is, and should be, a central phenomenon of interest to philosophy of mind, cognitive science, and psychology. Beyond its theoretical interest, moreover, it is arguably something of utmost importance to us in our personal lives, with direct and deep ethical implications.

Given this centrality of consciousness in our scheme of things, it is instructive to see how astonishingly varied and diverse the opinions are that influential thinkers have held about the nature and status of consciousness. We will begin with a sample of these views.

Some Views on Consciousness

It would be appropriate to begin with Descartes, who is often thought to have created the field of philosophy of mind:

> My essence consists solely in the fact that I am a thinking thing.[1]

For Descartes, being a thinking thing amounts to being a conscious being, as is made clear in his statement "There can be nothing in the mind, in so far as it is a thinking thing, of which it is not aware. . . . We cannot have any thought of which we are not aware at the very moment when it is in us."[2] For Descartes, then, my life is exactly coeval with my consciousness, which constitutes my essence; when I lose my capacity for consciousness, that is when I cease to exist. It is not surprising to be told that the loss of capacity for consciousness signifies the end of us as persons. But Descartes may be saying something stronger: Such a loss means our end as existing things. It isn't just that by losing consciousness we turn into something other than persons; we would simply cease to be—there is only nothingness beyond.

A similar sentiment is echoed by Ivan Pavlov, famed for his conditioning of dogs ("Pavlovian dogs") to salivate in response to a ringing bell, who avowed in his Nobel Prize acceptance speech in 1910:

> In point of fact, only one thing in life is of interest to us—our psychic life.[3]

This from a scientist whose work on conditioning was a major influence on the development of the behaviorist movement in psychology.

Views that are diametrically opposed have been expressed by some contemporary philosophers of mind. Daniel C. Dennett, well-known for his work on consciousness, achieved fame and, arguably, some notoriety, by boldly declaring:

> I want to make it just as uncomfortable for anyone to talk of qualia—or "raw feels" or "phenomenal properties" or "qualitative and intrinsic properties" or "the qualitative character" of experience—with the presumption that they, and everyone else, knows what they are talking about. . . . *Far better, tactically, to declare that there simply are no qualia at all.*[4]

Dennett does not deny the existence of all conscious states, but only those with intrinsic qualitative properties, or "qualia," like the painfulness of pains

and the green of a visual percept. Such a view is known as qualia nihilism or eliminativism.

Wilfrid Sellars, an eminent American philosopher a couple generations ahead of Dennett, responded:

> But Dan, qualia are what make life worth living![5]

This sounds like something Pavlov would have said. Experiences like seeing a glorious sunset over the glittering waves of the sea, smelling a blooming lavender field in a valley, and hearing a shifting, layered soundscape projected by a string quartet—these are among the things that make life worth living. On the other hand, we should not forget that qualia like pains from cluster headache, constant fears and anxieties, and unremitting depression and despair, may well be what makes life *not* worth living. The point, however, has been made: Conscious states are the source of all values for us, what is good and desirable and what is evil and to be avoided and deplored.

Dennett is not alone. Another philosopher of mind, Georges Rey, sounds intent on rejecting all forms of consciousness, not just qualitative consciousness:

> The most plausible theoretical accounts of human mentation presently available appear not to need, nor to support, many of the central claims about consciousness that we ordinarily maintain. One could take this as showing that we are simply mistaken about a phenomenon that is nevertheless real, or, depending upon how central these mistakes are, that consciousness may be no more real than the simple soul exorcised by Hume.[6]

The idea seems to be that consciousness has no role to play in a scientific account of human mentality and that in consequence it's wholly dispensable—its existence has no purpose to serve. We will have a chance below to discuss such a point of view (chapter 10).

Another theme that runs through many writings on consciousness is that consciousness is something mysterious and intractable to scientific study, and represents a serious hurdle to the understanding of how our minds work. The memorable remark by Thomas H. Huxley, a noted nineteenth-century English biologist, is a well-known example:

> But what consciousness is, we know not; and how it is that anything so remarkable as a state of consciousness comes about as the result of irritating

> nervous tissue, is just as unaccountable as the appearance of the Djin when Aladdin rubbed his lamp in the story, or as any other ultimate fact of nature.[7]

Huxley is joined by William James, regarded as a founder of modern scientific psychology, who wrote in his masterwork, *The Principles of Psychology* (1890):

> That brains should give rise to a knowing consciousness at all, this is the one mystery which returns, no matter of what sort the consciousness and of what sort the knowledge may be. Sensations, aware of mere qualities, involve the mystery as much as thoughts, aware of complex systems, involve it.[8]

The term "mystery of consciousness" has been frequently and freely bandied about; it is impossible to avoid it, especially in popular writings on mind, cognitive science, and neuroscience.[9] But the intractability of consciousness to an intelligible objective account is perhaps best expressed by Thomas Nagel in two short, pithy sentences:

> Without consciousness the mind-body problem would be much less interesting. With consciousness it seems hopeless.[10]

In a section to follow, we will discuss whether Nagel is right, especially about why he thinks it is hopeless to account for how consciousness relates to our bodily nature.

As you would expect, there are philosophers and scientists who take a positive and optimistic stance about the possibility of a scientific account of consciousness. Here is the Nobel Prize–winning molecular geneticist Francis Crick who turned later in his career to neural research on consciousness:

> Our approach [to consciousness] is essentially a scientific one. We believe that it is hopeless to try to solve the problem of consciousness by general philosophical arguments; what is needed are suggestions for new experiments that might throw light on these problems.[11]

Some people are attracted to the view that science will in good time unravel the mystery of consciousness just as it has done with the "mystery of life"; with advances in molecular biology we now understand how reproduction—the creation of life—is possible. It is quite common to see people take the attitude

"Who knows what our future science will accomplish? Just look at its past accomplishments. We should be patient and wait."

This attitude is nicely expressed by Patricia Churchland:

> The problems for neuroscience and experimental psychology are hard, but as we inch our way along and as new techniques increase noninvasive access to global brain processes in humans, intuitions change. What seems obvious to us was hot and surprising news only a generation earlier; what seems confounding to our imagination is routinely embraceable by the new cohort of graduate students. Who can tell with certainty whether or not all our questions about consciousness can eventually be answered?[12]

Churchland is convinced that the scientific approach is in principle capable of explaining consciousness in neural terms; as she indicates, it is an empirical question whether this will actually be accomplished. To philosophers, it is the first point, not the second, that is of primary interest.

Nagel and His Inscrutable Bats

In 1974, Thomas Nagel published a paper with the provocative title "What Is It Like to Be a Bat?" This landmark paper brought back consciousness from years of neglect and helped to restore it as a central problem in the philosophy and science of the mind.[13] Nagel accomplished this feat by vividly and forcefully arguing for the subjective character of conscious experience and its inaccessibility to an objective point of view, declaring, "With consciousness [the mind-body problem] seems hopeless." We can safely conjecture that there are no philosophers, or students of philosophy, with any interest in consciousness who have not read Nagel's paper or don't know Nagel and his bats.

One of the notable things in the paper is Nagel's definition of consciousness, which has gained quick currency and achieved a canonical status. This is the idea that to say that a creature has conscious experience means that there is "something it is like to *be* that creature." And there is a collateral idea: To say that a state of a creature is conscious is to say that there is "something it is like for the creature to be in that state." It seems correct to say: There is something it is like for us to experience pain in a burned finger, or to see a large red circle painted on a white wall, or to smell a rotten egg. So these states, like experiencing a pain, seeing a red circle, and so on, count as conscious states.

Nagel begins his argument by claiming that bats are conscious and that we have good reason to believe this. According to him, there is something it is

like to be a bat, and there surely must be something it is like for a bat to locate a flying moth by its echolocating sonar. There must be a representation of the moving moth in the bat's perceptual field, or so we are inclined to think. Actually, it seems like a curious, dubious idea that in addition, there is something it is like to *be a bat*. How would that differ from what it is like to be, say, an anteater? Is there something it is like to *be a human*? According to Nagel, there must be since we are conscious human beings. Can you locate, or identify, this humanlikeness that you experience? It seems that when you try to introspect deeply and carefully, you will not find anything like it; what you will find are the particular perceptions and mental states that are currently occurring. There is something it is like to see a tree, to experience an itch or pain, to feel uncomfortable sitting in a chair with a hard seat, and so on. These states and events are conscious, and you are a conscious creature because you can be in such states. If anyone should insist "I don't mean what it's like for you to experience an itch; what I want to know is what it's like for you to *be a human*?" it would be hard to know what to say, except perhaps "Compared with what?" Similarly, for the bats: We don't have to say there is something it is like to *be* a bat. Bats are conscious creatures because they are capable of conscious states, states such that there is something it is like for bats to be in them.

This small quibble aside, Nagel's term "what it is like" has become a widely used, almost standard way of explaining what it is for a state to be a conscious state—or, to be more exact, *phenomenally* conscious state. It is taken to single out the specific qualitative character of experiences, the redness of a visual percept, the hurtfulness of a pain, the smell of fresh newsprint, the tactile feel of a cool marble surface, and the like. These qualitative, or phenomenal, aspects of experiences are now generally referred to as "qualia" ("quale" for the singular), and qualia are at the heart of current debates on the nature of consciousness. More on this later.

At any rate, we can grant Nagel's two starting points: first, that bats are conscious creatures capable of having experiences with qualitative character, and that there probably are many other "alien" forms of consciousness; second, that we can know that bats are conscious without having any idea about the qualitative character of their consciousness—that is, without knowing what it is like to be a bat, or what it is like for a bat to echolocate a moth or to hang upside down in a dark cave. So the only answer to Nagel's question "What is it like to be a bat?" is that we have no idea. As Nagel puts it, bat experiences are beyond our conceptual reach—we have no conception of what they are like. Bat phenomenology—that is, bats' inner world of experience—is cognitively impenetrable to us.

Compare this with the cognitive and neural sciences of bats. Bats' cognitive abilities and their underlying neural mechanisms are pretty well understood; they can be scientifically investigated like any other natural phenomena. Details concerning how bats' echolocating sonar system works seem well-known; for example, bats emit ultrasonic sounds, at up to 130 decibels, and by comparing the emitted pulse with its echo, the bat brain constructs a detailed map of objects and events in its surroundings. It is known that some bats distinguish the outgoing pulse from its incoming echo by frequency, and others do this in terms of elapsed time. The sensitivity and reliability of echolocation can be tested in the expected ways—that is, observation of bats' activities under controlled conditions. We also have good knowledge of the power and limits of bats' visual system—their eyes and associated systems. We can find out what bats know about their environment and how they come to have such knowledge; we can also have a good idea of what they need and desire and what they do to achieve their aims. But that is not all there is about bats' lives: We are inclined to think that something more is going on with echolocating bats, over and beyond these neural-behavioral processes. We are tempted to think that when a bat echolocates a moth flitting across its perceptual field, the bat must have an experience with a certain character, and that there must be some inner representation of the moving moth in the bat's "mind." And that is exactly what we are missing, something apparently beyond the reach of the neural-behavioral science of bats. So the following seems what we can say about Nagel's bats at this point:

> We can know all about bats' behavior and physiology but nothing about the qualitative character of their experiences.[14] An ideally complete neurophysiology, cognitive science, and behavioral psychology of the bats will tell us nothing about the phenomenology of bats' experience.

One worry that this view is likely to provoke is whether it leads to general solipsism, the consequence that not only can you not know what it's like to be a bat and experience bat pains, you cannot know what it's like to be another person or what it's like for another person to experience pain and joy. Aren't we cut off from the inner lives of other humans as much as from bats' inner worlds? Nagel denies this implication; he claims that we can, and do, know what it is like to be another human person, and what it is like for her to have experiences of various kinds. What then is the difference between other persons and bats? You can no more experience another person's pain, "from the inside," than a bat's pain. Nagel's response is that there is the following crucial

difference: While you can take "the point of view" of another human person, you are not able to take a bat's point of view. Talk of "points of view" occurs prominently throughout Nagel's discussion of consciousness, and it seems intimately tied to his notions of consciousness and subjectivity. But what is it for me to "take a point of view"—mine, or yours, or a bat's? Is it really anything more than a metaphor? If I can take your point of view but not a bat's, what is it that I try to do and succeed in your case but fail in the case of the bat?

We do not get much guidance from Nagel on this question. References to "points of view" or "perspectives," as in "first-person point of view" and "third-person perspective," are frequently encountered in philosophical discussion of consciousness and subjectivity,[15] but these expressions are never explained with acceptable clarity. The only firm idea in this area seems to be that every experience has a subject, a single unique subject whose experience it is, and this is taken to mean that the experience is experienced from that single subject's point of view. Further, it is a "first-person" point of view in that it is the experiencer's point of view. Your point of view, in regard to an experience of mine, is a third-person point of view, and this means only that it is not a first-person point of view, that is, you are not an experiencer of my pain.

On this explanation, there would appear to be no difference, for me, between your experiences and a bat's experiences. I am a third person with respect to both. So, again, what is the difference? Why don't Nagel's views about bats lead to general solipsism? Nagel appears to have in mind something like this: I can *empathetically* take your first-person point of view, and see the world, or imagine seeing the world, as you do. But empathetic identification is what is not possible with bats; I simply cannot imagine, or conceive, seeing the world the way bats experience it. This sort of explanation, which includes talk of empathetic identification, imagining oneself in another subject's mind, and so on, may have some explanatory value but it hardly makes anything clearer or more precise; if anything, it only introduces more complications.[16] It leaves unanswered the critical question how imagining in this sense can yield *knowledge* of another mind. How do I know whether I am imagining correctly, not just fantasizing or making things up? In imagining how your experiences seem to you, why am I not merely superimposing my experiences on yours, making this so-called imagination a case of reading my own mind?

We must set aside these questions and move on. For there is another question, a more urgent one, that could make these questions about imagination and knowledge of other minds immaterial as far as Nagel's argument is concerned. Suppose, *per impossibile*, that we could somehow peek inside a bat's mind and find out what it's like to be a bat. Would that help us to derive, or

otherwise acquire, knowledge of bat phenomenology from bat physiology, or to build an objective physical account of bat consciousness? The answer has to be a clear no. If we knew what it's like to be a bat, that might show us what we need to derive from bat physiology, or what we must build an objective account of, but we can be sure this will give us no material help. A superintelligent bat neuroscientist could no more derive bat phenomenology from bat physiology than we can. After all, we know what it's like to be a human and are familiar with the sorts of experience that we are capable of as humans. But that does not help us one bit with a derivation of facts about our consciousness from facts about our brains. Would an ideally complete neurophysiology of human brains give us knowledge of the phenomenology of human conscious experiences? Whatever the correct answer is to this question, it clearly does not depend on who, other than us, has cognitive access to what it's like to be a human being.

What this shows is that the impression that the cognitive impenetrability of bat consciousness is a crucial premise in Nagel's argument has to be mistaken—or at best misleading. Even if this premise, heralded by the attention-catching title of Nagel's paper, were false, or simply set aside, the physical irreducibility of bat consciousness could still stand; whether bat consciousness is accessible to us is irrelevant to the conclusion. In short, the bat story plays no role of significance in arguments about the physical reducibility of consciousness. It is best viewed as a vivid, and interest-provoking, preamble—a "consciousness raiser," if you like!

What then is Nagel's argument for the irreducibility of consciousness? The following line of reasoning can be discerned: Consciousness is essentially subjective in the sense that conscious states are accessible to a single unique subject (that is, from a single "point of view"), whereas physical states, including the states of the brain, lack this subjective character. A reduction of consciousness to brain states would turn essentially subjective states into states without subjectivity—that is, conscious states have been turned into states that are not conscious. But that is absurd, and hence there can be no consciousness-to-brain reduction. This line of argument invites many questions, but it is an intelligible, and not obviously implausible, line of thought.

Phenomenal Consciousness and Access Consciousness

At this period, it will be useful to do some initial clarification of terminology. "Conscious," as an adjective, applies both to persons and organisms and to

their states. We are conscious creatures, but cabbages, amoebas, and flowerpots are not. This does not mean that we are conscious at every moment of our existence; we are conscious when we are awake and alert and not conscious while in deep sleep or a coma. We are also conscious, or aware, *of* things, states, or facts (for example, the blinking red traffic lights, a nagging pain in the back, being tailed by a police patrol car). In cases of this sort, "conscious" and "aware" have roughly the same meaning. Moreover, we also apply the term "conscious" to events, states, and processes. Familiar sensory states and events, such as pains, itches, and mental images, are conscious states and events. Emotions, like anger, joy, and sadness, are usually conscious, although it is common to acknowledge unconscious desires, resentments, and so on. Many of our beliefs, desires, hopes, and memories are conscious, but there are many others of which we are not consciously aware. I have a conscious belief that I should send a contribution to my alma mater's annual fund, and I look for my checkbook; an observer, however, might say that all that is caused by my unconscious desire not to disappoint my friend, the class agent.

What is the relationship between these two uses of "conscious," as applied to persons and creatures ("subject consciousness," we may call it) and as applied to their states ("state consciousness")? Can we say something like this: "A state of a person is a conscious state just in case that person is aware, or conscious, of it?" It is plausible that if a state is a conscious state, the person whose state it is must be conscious of it. The converse, however, seems false: You are aware of your age and weight but that doesn't make your age and weight conscious states. You are aware of your posture and orientation (whether you are standing or lying down), and this kind of proprioceptive awareness seems direct and immediate, but these bodily states are not conscious states. The suggested relationship between state consciousness and subject consciousness makes better sense when restricted to mental states: A mental state is a conscious state just in case the subject whose state it is is conscious of it. (We will further discuss this idea in connection with "higher-order" theories of consciousness below.) The converse direction, from conscious state to conscious creature, seems direct and simpler: We can explain a conscious creature as one that is capable of having conscious states. In this and the following chapter, we will be mainly concerned with state consciousness—that is, the nature and status of conscious states—and not much with subject consciousness.

It is now customary to distinguish between two types of consciousness, "phenomenal consciousness" and "access consciousness."[17] This distinction is important because the two types are thought to present different sets of philosophical issues, and an account of one type of consciousness does not auto-

matically transfer to the other type. And when we are offered an "account" or "theory" of consciousness, we need to be clear about what sort of consciousness it is an account or theory of.

Phenomenal Consciousness: "Qualia"

When you look at a ripe tomato, its color looks to you a certain way, a way that is different from the way the color of a mound of lettuce leaves looks. If you crush the tomato and bring it close to your nose, it smells in a distinctive way that is different from the way a crushed lemon smells. Across sense modalities, smelling gasoline is very different from tasting it (or so we may safely assume!). Sensory mental events and states, like seeing a ripe red tomato, smelling gasoline, experiencing a shooting pain up and down your leg, and the like, have distinctive qualitative characters, that is, felt or sensed qualities, by means of which they are identified as sensations of a certain type. It is standard to refer to these sensory qualities of mental states as "phenomenal" (sometimes "phenomenological") properties, "raw feels," or "qualia."[18] Conscious states with such qualitative aspects are called phenomenal states; they are instances of phenomenal consciousness. Perceptual experiences and bodily sensations are among the paradigmatic cases of this form of consciousness.

Another standard way of explaining the notion of phenomenal consciousness is to resort to the idiom of "what it's like," which is already familiar to us from Nagel's "Bat" paper. A phenomenally conscious state is a state such that there is something it is like to be in that state. For example, Ned Block writes:

> Phenomenal consciousness is experience; what makes a state phenomenally conscious is that there is something "it is like" to be in that state.[19]

"What it's like" is supposed to capture the qualitative character, the quale, experienced in an experience. I know what it is like to see a golden yellow patch against a dark green background (when I am looking at a Van Gogh landscape), but a person who is yellow-green color-blind presumably does not know what it is like *for me* to have this experience. Conversely, normally sighted persons do not know, at least firsthand, what it is like for the color-blind person to see yellow against green.

Block's characterization of phenomenal consciousness, if taken literally, appears too broad. There is something it is like to hold your breath for one full minute, or to meet the president in the Oval Office. But meeting the president and holding your breath are not conscious states, phenomenally or otherwise.

Rather, it is the *experience* of meeting the president, or of holding your breath for a prolonged time, that is conscious. As Block says in the same quotation, phenomenal consciousness is experience. Even with this caveat, however, it is not clear that the popular locution "what it's like" is fully adequate to pick out the qualitative character of an experience. Consider pain: There surely is something it is like to have a sharp, stabbing pain in your elbow. But, as Christopher Hill observes, what it's like may include too much:[20] Besides the felt quality of painfulness, what it's like to be in pain may also include a feeling of anxiety and a desire to be rid of it, an awareness of your bent and elevated elbow, and the like. This can vary from one pain to the next, and from person to person. What it's like may be too inclusive in another way: It seems correct to say there is something it's like to believe something, as distinguished from doubting or being unsure. But is there a belief quale, a special qualitative character to all beliefs? Or consider being aware. There surely is something it is like to be awake and aware. But is there a quale of awareness, a qualitative character that attaches to awareness as such? The content of awareness at a given time may be constituted by various qualia, like colors, shapes, and sounds. But it's questionable whether there is an awareness quale, or a belief quale, in the sense in which there is a pain quale. It seems then that the idea of what it's like and the idea of qualitative character of experience do not fully coincide. What it's like appears to define a broader class than phenomenal character.

Emotions in general have qualitative aspects. Anger, remorse, envy, pride, and other emotions appear to have distinctive qualitative feels to them; after all, emotions are "experienced" and often intensely "felt." Unlike sensory experiences, however, they do not seem to be type-classified solely, or primarily, on the basis of how they feel. For example, it may be difficult, perhaps impossible, to categorize an emotion as one of resentment, envy, or jealousy based on its felt qualities alone. Nor does every instance of an emotion need to be accompanied by a distinctive felt character. Suppose you are unhappy, even upset, about the continuing large budget deficits of the federal government. Must your unhappiness be accompanied by some special felt quality? Probably not. Even if it is, must the same quality be present in all other cases in which you are upset or unhappy about something? Being in such a state seems more a matter of having certain beliefs, attitudes, and dispositions (such as the belief that large budget deficits are bad for the economy, or your eagerness to work for the opposition in the next election) than having an experience with a distinctive felt quality. Moreover, it is now commonplace to acknowledge emotions of which the subject is not aware (for example, repressed anger and

resentment), and it seems that such unconscious states cannot be constituted, even in part, by phenomenal qualities.

Moods are often classified with emotions. There certainly are distinctive felt qualities to moods, like "good" and "bad" moods, mildly depressive moods, and positive, upbeat moods. In some respects, moods seem more like bodily sensations than emotions; for example, moods seem type-classified primarily, if not exclusively, in terms of their qualitative character: A bad mood feels a certain way and a good mood feels another way. As we saw, emotions can be unconscious, but it seems at best odd to speak of unconscious moods. And moods are not focused in the way emotions typically are.

Returning to beliefs, might it be the case that all beliefs share a special distinctive phenomenal feel? The answer must be no—for the simple reason that, like unconscious emotions, there can be, and are, unconscious beliefs, and it seems at least odd to associate a phenomenal quality with mental states of which we are unaware. Freudian depth psychology, parts of which seem to have been assimilated by commonsense psychology, has told us about the psychological mechanism by which we repress beliefs, desires, and emotions that are unacceptable to our conscious minds. The media often carry news of people whose repressed memories of childhood abuse are recovered through therapy. But we do not need such controversial cases. If I ask you, "Do you believe that some neurosurgeons wear hats?" you would probably say yes; you do believe that some neurosurgeons wear hats. This is a belief you have always had (you would have said yes if I had asked you the question two years ago); it is not a new belief that you have just acquired, although you became aware of it only now. My question made an unconscious belief an "occurrent" one. Obviously, there are countless other such beliefs that you have. These beliefs are called "dispositional" beliefs. Dispositional beliefs are not experienced, and if we follow Block's statement that phenomenal consciousness is experience, we have to say that there is nothing it's like to have a dispositional belief and that it has no phenomenal character.

What then of conscious beliefs? Are all conscious instances of the belief that George Washington was the first president of the United States, in different persons or for the same person at different times, characterized by a special qualitative character unique to beliefs with this content? The answer must be no: One person with this belief might have a mental image of George Washington (as on the dollar bill), another may simply have the words "George Washington" hovering in her mind, and still another may have no particular mental image or any other sort of phenomenal occurrence. There is also the

general question: Do all occurrent beliefs—beliefs we are actively entertaining—share some specific belief-like phenomenal character, a belief quale? Some have claimed that when we are aware of a thought as a belief, there is a certain feel of assertoric or affirmative judging, a sort of "Oh, yes!" feeling. Similarly, we might claim that an occurrent disbelief is accompanied by a directly experienced feel of denial and that remembering is accompanied by a certain feeling of déjà vu. Perhaps desiring, hoping, and wishing are always accompanied by a felt feeling of yearning or longing combined with a sense of present privation.

But such claims are difficult to assess. The "Oh, yes!" feeling may be nothing more than our coming to be aware that we believe a certain proposition. Such awareness does not seem to be accompanied by a particular kind of felt quality. When your physician wants to know how the pain in your bruised elbow is coming along, you can focus your attention on the elbow and try to see whether the pain is still there. Here there clearly is a special kind of sensory feel, a quale, that you are looking for. However, when you are unsure whether you really believe some proposition—say, that euthanasia is morally permissible, that Velasquez is the greatest painter in Western history, or that the Republicans will make a comeback in the fall of 2010—you do not look for a quale of a special type.[21] It is absurd to suggest that there is some sensed quality such that if you find it attached to your thought that euthanasia is morally permissible, you say, "Aha! Now I know I believe that euthanasia is morally permissible," and if you don't find it, you exclaim, "Now I know—I don't have that belief!"

If there are no distinctive phenomenal qualities associated with types of intentional mental states—beliefs, desires, intentions, and the rest—we face the following question: How do you know that you *believe*, rather than, say, merely *hope*, that it will rain tomorrow? Such knowledge, at least in normal circumstances, seems direct and privileged: it is not based on evidence and it would be highly unusual, perhaps incoherent, to think you could be mistaken about such matters. One thing that is certain is that we do not find out whether we believe or hope by looking inward to detect specific qualia. Nor is it obvious that we know that we are angry, or that we are embarrassed, by detecting a special phenomenal quality. How, then, do we know that we are angry rather than embarrassed? Or embarrassed rather than ashamed? Sometimes we find that it is not possible to classify our feeling as one of embarrassment or shame—perhaps it is both. But then how do we know *that*?

One possible answer might go like this: I know what it's like to be embarrassed and what it's like to be ashamed; that's how I know I am embarrassed, not ashamed. But what is the nature of such knowledge? What do you know

when you know what it's like to be angry? Perhaps it's like knowing what apples taste like and knowing that apples and oranges don't taste alike. If so, something like phenomenal knowledge might be involved in the knowledge of our own intentional states—whether they are beliefs, desires, hopes, and the like. Again, what this shows is that the notion of a quale, or phenomenal character and the notion of what it's like, can come apart. Though it seems to make no sense to speak of, say, belief-like quale, it seems true that there is something it is like to believe (rather than to doubt), and that we know what it's like to believe. In any case, it is plausible that there are conscious mental states with no special phenomenal qualities, though there is something it's like to be in them. Mental occurrences that we call "experiences" appear to be those that possess phenomenal properties. Sensing and perceiving are experiences, but we do not think of believing and thinking as experiences.

To sum up, mental states come in two groups—those of which the subject is conscious or aware and those of which the subject is not aware or conscious. We can leave it open whether there are conscious states that are not accompanied with qualia, or that do not have a phenomenal character. It seems that in the sense of what it's like, all conscious states may have a phenomenology: There is always something it is like to be in a conscious state, although we have raised questions as to whether a characteristic quale can always be associated with conscious states. In any case, conscious states with qualitative characters can be divided into two subgroups: those that are type-classified or type-individuated on the basis of their qualitative character (for example, pain, smell of ammonia, visual sensing of green) and those, even though they are accompanied by qualia, that are not type-classified in terms of their qualitative character (such as beliefs, doubts, and emotions). Perhaps we can mark this last division by saying that conscious states in the first group are wholly *constituted* by qualia and that those in the second group, though they possess qualitative characters, are not constituted by them, perhaps not even partially.[22]

Access Consciousness

You happen to look out your window, interrupting your work, and see a heavy rain splashing your window and water gushing down the gutters of the street in front of your building. You become conscious, or aware, that a heavy rain is coming down in your area. As a result, you decide to take an umbrella with you on your way out to lunch; you also call your friend who is driving to Boston this afternoon. You tell her, "It's raining very hard. So be extra careful on the highway."

Now, if you were not consciously aware of the rain (you were completely absorbed in your work, unaware of what was going on outside), you would have done none of these things. The point to note is that the content carried by your conscious state, in virtue of your state being conscious, has become available to various other cognitive functions, like reasoning (your inference that driving may be difficult in the afternoon), decision making (your deciding to take an umbrella), and verbal reporting (your telling your friend about the rain). That is, these cognitive faculties or modules have *access* to your conscious state about the rain. That is the basic idea of access consciousness. Ned Block, who formally introduced the notion, writes:

> A state is A-conscious [access-conscious] if it is poised for direct control of thought and action. To add more detail, a representation is [access]-conscious if it is poised for free use in reasoning and for direct "rational" control of action, and speech. An [access-conscious] state is one that consists in having an [access-conscious] representation.[23]

The point of adding "rational" to control of action can be seen by considering conscious and unconscious desires. When you have a conscious desire, for example, to go to medical school, it can enter into your rational deliberation on actions and decisions. On the other hand, an unconscious desire, say, a desire to outshine your siblings in your parents' eyes, can influence and shape your behavior and actions but will do so only causally, not through rational planning and deliberation. Its content is not freely available for rational decision making or verbal reports. Similar examples are possible about conscious and unconscious beliefs.

It is clear that only those mental states with representational content can be access-conscious. So, moods, for example, seem to fail to be access-conscious in this sense since they are not representations and don't have representational content. You might ask why. The short answer is that a genuine representation must have "satisfaction conditions," conditions that define its representational correctness, accuracy, or fidelity. Thus, your consciousness of the rain is a representational state because its representation may be correct or incorrect; it is correct if it is raining and not so if it isn't. In contrast, moods don't have satisfaction conditions; they cannot be accurate or inaccurate, true or false. Are bodily sensations, like pain and itch, representational? Do they have satisfaction conditions? We turn to this question later.

Various theories of consciousness deal with access consciousness. Bernard Baars, a cognitive neuroscientist, has proposed what he calls "the global work-

place theory," a theory of consciousness at the cognitive level (as distinguished from the neural level).[24] The central idea is that a mind is a kind of theater, a "global workplace," in which conscious states "broadcast" themselves so that their representational contents are available to various other cognitive functions and processes. Evidently the conception of consciousness involved is Block's access consciousness.

Another theory in this vein is Daniel Dennett's "multiple draft" theory.[25] Dennett conceives of our perceptual-cognitive system to be constructing multiple pictures ("drafts") of our surroundings. The draft that gains prominence at a time, the one that, in Dennett's term, achieves "cerebral celebrity," is our conscious state at that time. It is this state whose representational content has the largest influence on our cognitive systems.

Many, perhaps most, conscious states are both phenomenally conscious and access-conscious. Perceptual experiences are normally both access-conscious and phenomenally conscious. But if access and phenomenal consciousness are really distinct forms of consciousness, there should be cases, at least possible cases, of conscious states that are phenomenally conscious but not access-conscious, and vice versa. Are there such states? Most of us have had an experience of suddenly "coming to" while driving and then realizing that we don't remember anything about the road or traffic during the past half hour. What happened during the half hour? Your visual perception was active—otherwise you would have run off the road. You surely had visual percepts with their phenomenal qualia. On the other hand, you were not actively aware of what was going on, and the representational content of your consciousness was not available for verbal reports or short-term memory. Though it did have an influence on your driving, this was purely causal and automatic (you were on "automatic pilot," after all), it had no role in the rational guidance of your behavior or practical reasoning. Arguably this is a case of phenomenal consciousness without access consciousness.

It is a little harder to think of a case where we have access consciousness without phenomenal consciousness. The reason may be that when you are in an access-conscious state, it seems that you must be aware and awake, and there surely is something it is like to be awake and aware. If we understand phenomenal consciousness in terms of qualia, or phenomenal properties, being aware may not count as a phenomenally conscious state since, as we observed earlier, there seems to be no such thing as awareness quale. And it is possible to find potential cases of awareness without phenomenal character. One pretty plausible case of access consciousness without phenomenal consciousness is Block's "super blindsighter."[26] (Blindsight is a syndrome observed

in patients with damage to their primary visual cortex who as a result have blind areas in their visual field, where they report no visual percepts. However, when stimuli are flashed in the blind field, a patient is often able to guess whether it is, say, an "O" or an "X," and its location and motion, and is even able to catch a ball thrown in the blind field.)[27] A blindsight patient has been trained to guess, on her own without anyone's prompting, whether there is an "X" or an "O" in her blind field, and becomes aware, say, that she is presented with an "X." This awareness is access-conscious since its content is now available for the use of her cognitive functions, but there is no visual phenomenology in this awareness—it lacks phenomenal consciousness.

Finally, one important point: Access consciousness is a functional concept. A mental state is access-conscious just in case it performs a certain function in our cognitive economy: Its representational content is freely available for various other cognitive functions and activities, such as reasoning, rational guidance of behavior, short-term memory, verbal reports, and so on. This makes access consciousness a proper subject of investigation in cognitive and neural science, and we can expect various theories couched in information-processing terms, that is, computational models of how access consciousness works. Phenomenal consciousness, in contrast, is not characterized in terms of its function but in what it is like in itself, that is, in terms of its intrinsic properties. As we will see shortly, there is the view that phenomenally conscious states, too, are essentially representational, and that qualia can be fully explained in terms of their representational properties. This is consciousness representationalism. If it is true, all consciousness is at bottom representational and functional in nature. This is one of the central issues being debated in the field. But it remains true that phenomenal consciousness is not initially defined in terms of any function it is to perform; it is not a functional concept.

Consciousness and Subjectivity

Conscious mental states are often taken to have certain special properties. Here we canvass several of these.

Subjectivity and First-Person Authority

As we saw with Nagel and his bats, subjectivity is often claimed to be of the essence of consciousness.[28] However, subjectivity has no fixed, unambiguous meaning. One sense of subjectivity is epistemological, having to do with the

supposed special nature of knowledge of conscious states. The main idea is that a subject has a special epistemic access to her own current conscious states; we seem to be "immediately aware," as Descartes said, of our own feelings, thoughts, and perceptions and enjoy a special sort of first-person authority with regard to them.

A precise explanation of just what this "immediate" or "direct" access consists in is a controversial question despite the virtually universal consensus that special first-person epistemic authority is real.[29] However, the following three features are notable, and we will try to put them in untendentious terms: (1) Such knowledge is not based on, or inferred from, evidence about other things—observation of what we say or do, what others tell us, physical or physiological cues, and the like. Your knowledge that you are having a toothache, that you are thinking about what to do this weekend, and such is *direct* and *immediate* in that it is not based on other things you know. (2) Your knowledge of your own current mental states carries special authority—"first-person authority"—in the sense that your claim to have such knowledge cannot, except in special circumstances, be overridden by third-person testimony. Concerning the queasy feelings you are experiencing in your stomach, your afterimages, the itchy spot on your shoulder, what you are thinking about, and other such matters, *what you say goes, at least in normal circumstances*, and others must defer to your avowals. As the qualifying proviso implies, first-person authority need not be thought to be absolute and unconditional. There is psychological evidence that seems to show that we do make mistakes about what beliefs we hold.[30] In any case, whatever the degree or firmness of first-person authority, it is evident that the subject does occupy a special position in regard to her own mind. We should keep in mind the possibility, in fact the likelihood, that our access to intentional states, like beliefs and desires, may be different in nature from our knowledge of phenomenal states, like pains and perceptual sensations. Note, too, that points (1) and (2) hold only for a subject's *current* mental states, not her past or future ones.

Finally, (3) there is an asymmetry between first-and third-person knowledge of conscious states. Neither of the foregoing two points applies to third-person knowledge, that is, knowledge of another person's conscious states. The subject alone enjoys immediate and special authoritative access to her conscious states; the rest of us have to listen to what she has to say or observe her behavior or examine her brain. The idea that minds are "private" reflects this epistemic asymmetry between first-and third-person access to mental states.

Experience and the First-Person Point of View

As we saw in connection with Nagel and his bats, some philosophers closely associate subjectivity of consciousness with the notion of a first-person *point of view*, or *perspective*. Nagel writes:

> If physicalism is to be defended, the phenomenological features must themselves be given a physical account. But when we examine their subjective character it seems that such a result is impossible. The reason is that *every subjective phenomenon is essentially connected with a single point of view*, and it seems inevitable that an objective, physical theory will abandon that point of view.[31]

And in a later essay:

> *Yet subjective aspects of the mental can be apprehended only from the point of view of the creature itself* . . . , whereas what is physical is simply there, and can be externally apprehended from more than one point of view.[32]

Nagel is saying that the subjectivity of mental phenomena is essentially connected with—perhaps consists in—there being "a single point of view" from which these phenomena are apprehended. This fits in with the idea that conscious states have phenomenal features in the sense that there is something it is like to be in such states. For there can be no impersonal "what it is like"; it is always what it is like *for a given subject* (for you, for humans, for bats) to see yellow, to taste pineapple, to locate a moth in flight. Things do not look, or appear, this way or that way, period; they look a certain way to one perceiving subject and perhaps a different way to another. There are no "looks" or "appearances" or "what it's like" in a world devoid of subjects capable of having experiences.[33]

Understood this way, talk of "points of view" appears to have two parts: First, as noted earlier, there is the idea that for any conscious state there is a conscious subject whose state it is, and that the content of consciousness consists in how things look or appear to that subject. Second, the subject knows, or "apprehends," the content of the conscious state in a way that is not open to anyone else. Here is the important point: This isn't merely a matter of the subject's greater authority, or the higher degree of certainty he can achieve; nor is it just a matter of his being more reliable as a witness to what goes on in his mind. It isn't a matter of degree in epistemic access, authority, or reliability.

The nature, or kind, of access seems qualitatively different in the first-person and the third-person cases. And this difference might explain the special epistemic authority that attaches to the knowledge of one's own conscious state. What could this difference be?

The only answer has to be that what is special and different in my relation to my pains, when contrasted with your relation to my pains, is that I *experience* my pains and you do not. I am the experiencer and you are an observer. To say simply that I am the thing that has the pains, or that instantiates pain, like I am the thing that instantiates the property of being right-handed or being brown-eyed, doesn't quite capture it. I experience my experiences and that is how I get to know what it is like to have, or undergo, them—what it's like to hurt, to smell burned bacon, to see green. There is also the corollary idea: We get to know what an experience is like only by experiencing it. This fits in with the idea, at least as old as Locke, that phenomenal properties, like the taste of pineapple and the smell of lavender, can be known or grasped only by experiencing them—you have to taste pineapple and smell lavender to grasp them. Christopher Hill writes:

> One approach is to say that qualia are properties that we normally think of as *subjective*, in the sense that it is possible to grasp them fully only from the point of view of an experiencing subject.[34]

Actually, even "grasping" doesn't seem strong enough, or fully apt. Your knowledge that I am in pain may have as much certainty and warrant as my knowledge that I am in pain. But I am hurting and you are not! In this sense, experiencing isn't simply a matter of having knowledge of some superior kind, or even "grasping" what it is like. Pains, and other experiences, matter to us because we experience them, and this is not a mere epistemic point. The kind of experiential intimacy this suggests seems to go well beyond what can be captured in epistemic-cognitive terms. There are issues in this area that deserve further thought and reflection, but we must move on.

Does Consciousness Involve Higher-Order Perception or Thought?

One idea that has been influential in discussions of consciousness is that consciousness involves a kind of inner awareness—that is, awareness of one's own mental states. The model is that of a kind of scanner or monitor that keeps tabs on the internal goings-on of a system. Consider again the experience of driving

on "automatic pilot": You perceive the conditions of the road and traffic, but there is a sense in which your perceptions are not fully conscious. That is, you are not aware of what you see and hear, although you do see and hear, and you are unable to recall much of anything about the conditions of the traffic for several minutes at a time. Or consider pains: In the heat of competition or combat, an injured athlete or wounded soldier can be entirely unaware of pain. His attention is wholly occupied with other tasks, and he is not conscious of pain. In such a case we may have an instance of pain that is not a conscious pain, and the reason may be that there is no awareness, or internal scanning, of the pain.[35]

David Armstrong has advocated an account of this kind. According to Armstrong, consciousness can be thought of as "perception or awareness of the state of our own mind."[36] Return to the absent-minded driver: The driver perceives the conditions around her and makes automatic adjustments to keep the car going in the right direction at the right speed, but she has no awareness of her perceptions. The injured athlete does not perceive his pain, and this is exactly what makes his condition nonconscious. When he comes to notice the pain, it becomes a conscious pain. So there are "first-order" perceptions and sensations (you seeing another car moving past you on your left, the pain in your knee), and there are perceptions of these first-order perceptions and sensations—that is, "second-order" or "higher-order" perceptions. We can then say something like this: A mental state is a conscious state just in case there is a higher-order perception of it—or perception of being in that state. And a creature is conscious just in case it is capable of having these higher-order perceptions. An approach like this is called a "higher-order perception" theory (or HOP) of consciousness.

In a similar vein, David Rosenthal has proposed that "a mental state's being conscious consists in one's having a *thought* that one is in that very mental state."[37] On this account, then, a mental state is a conscious state just in case there is a higher-order thought, or awareness, that one is in that state. As you would expect, this approach is called the "higher-order thought" theory (or HOT). So consciousness involves a sort of "metapsychological" state, that is, a psychological state about another psychological state. A view like this typically allows the existence of mental states, even sensory states, that are not conscious—that is, those not accompanied by higher-order thoughts. And this for good reason—otherwise there would be an infinite progression of higher and still higher mental states, without end.

How plausible is this view of consciousness? There is no question that it has a certain initial plausibility and that it nicely fits some typical cases of mental states that we recognize as conscious. In particular, it seems to make

good sense of Block's notion of access-consciousness: Our perception, or thought about, a mental state makes the content of that state accessible to other cognitive activities, like verbal reports and decision making. And the Armstrong-style view seems to open the door to a functionalist explanation of consciousness: On the functionalist view, first-order perceptions receive an account in terms of their causal roles or functions—in terms of their typical stimulus conditions and behavioral-psychological outputs—and if this is right, we might plausibly attempt a similar functional account of consciousness as an internal monitoring activity directed at these first-order perceptions and other mental states. Perhaps such an account could explain the role of consciousness in organizing and coordinating disparate perceptions—perceptions coming through different sensory channels—and even yield a functional account of "the unity of consciousness."

At every moment of our waking lives we are bombarded by sensory stimuli of all sorts. The role of consciousness in the proper coordination and integration of an organism's myriad sensations and perceptions and the selection of some of them for special attention may be crucial to its ability to cope with the constantly changing forces of its environment, and this means that an approach such as Armstrong's may fit well with an evolutionary explanation of the emergence of consciousness in higher organisms. This indicates that higher-order theories are well-positioned to give an account of access consciousness and make it amenable to computational modeling in cognitive science. When there is a higher-order awareness of a mental state, we can expect the content of the state to be made available for various cognitive-executive systems. There is a general agreement on that point. The main issue with the higher-order approach is whether it can deliver a satisfactory account of phenomenal consciousness, the "what it's like" aspect of consciousness.

There are various forms of higher-order theories—we have seen two above, the higher-order perception and higher-order thought theories, and each of these has variant versions. All these theories give divergent accounts of phenomenal consciousness, but there are commonalities.

According to the higher-order perception approach, your pain, for example, is a conscious pain just in case you perceive that pain. Your pain, the first-order mental state, is supposed to have "nonconceptual" content, and your higher-order perception of the pain too is supposed to be a nonconceptual state. Conceptual content is content that can be expressed linguistically, in terms of concepts and sentences. Thus, propositional attitudes, like beliefs, have conceptual contents represented by declarative sentences; the contents are propositions composed of concepts. In contrast, nonconceptual content is

not represented linguistically, or in terms of concepts; they are like pictures and maps and can be visual, tactile, auditory, and so on. Of course, you can read off conceptual contents from your visual percepts, but visual percepts will in general be far richer in content than what can be captured conceptually. (You can discriminate far more shades of red than you have concepts or terms to refer to each.) For this reason, nonconceptual content is said to be "fine-grained" or "analog."

The higher-order thought account of phenomenal consciousness would go like this: A mental state is phenomenally conscious just in case it is a state with nonconceptual content and there is a higher-order thought, or awareness, that one is in that state. The difference between the two higher-order theories consists in this: According to the higher-order thought theory, the second-order mental state has conceptual content—it's discursive and has the form "I am in a state of type P" or "a state of type P is occurring"—whereas on the higher-order perception approach, the second-order state itself is a nonconceptual perceptual state.

Let us first consider the higher-order perception theory. Various questions can be raised about the second-order, or inner, perceptual mechanism that is supposed to perceptually scan the first-order mental event. Suppose the first-order state is one of seeing green. The second-order perception is supposed to have that first-order state as its target and itself have a nonconceptual content. What could this nonconceptual content be? When I perceive my perception of green, what do I perceive? It is hard to say what it might be—unless the second-order perception "inherits" its content from the first-order perception, namely green. There is the influential view that perceptual experiences are "transparent" or "diaphanous." You are looking at a round green spot on the wall. Try to focus on your visual experience of the green spot; you will soon realize that all that happens is that you see "right through" your visual experience and end up focusing on the green spot on the wall.[38] (We will discuss this supposed phenomenon below in connection with qualia representationalism.) But doesn't this mean that the second-order perception simply collapses into the first-order perception? Your supposed second-order perception of your first-order perception of a green spot may turn out to have the same nonconceptual content as its target.

Second, the talk of scanning, or monitoring, first-order mental states cannot be just a metaphor. If second-order perception is real, there must be a physical-neural organ that does the scanning and monitoring. If there is such an organ, it could malfunction and give us incorrect reports about first-order states; for example, the first-order perception is of a green spot but the second-

order perception of that perception reports it's red, or it's the smell of ammonia. Can this happen? Does it make sense? And do we have any empirical evidence that there is a neural system that does the second-order perceptions?

Let us move on to the higher-order thought account of phenomenal states. One apparent problem with this approach is that, as we noted briefly above, phenomenal states, or qualitative contents or qualia, seem to far outrun our available concepts. Suppose Q_1 and Q_2 are two distinguishable color qualia, both shades of red. On this account of qualia, to say that these are distinct qualia seems to imply that the second-order thought that Q_1 is instantiated is different from the second-order thought that Q_2 is instantiated. And these two second-order thoughts are different only if Q_1 and Q_2 are represented by distinct concepts. It apparently follows that for each quale I experience, I must have at my disposal a distinct concept for it—that is, I must have as many concepts as all the qualia that I experience, or can experience. But that can't be true; it is an established psychological fact that most of us can distinguish far more shades of red, or far more hues, than we have concepts or words to designate them.

A second difficulty concerns a possible infinite regress that threatens the higher-order thought theory (this objection may apply also to the higher-order perception theory). It might be argued that the second-order thought about a first-order mental event might itself be unconscious. Higher-order theorists in general accept the possibility that any mental state could occur as an unconscious state. But if the second-order thought is unconscious, how could that make the first-order state conscious, phenomenally or otherwise? Two unconscious states could be such that one of them is about the other; but how could a conscious state emerge from this pair? The only clear escape seems to be the requirement that the second-order thought be a conscious state. But to make the second-order state conscious, we need a conscious third-order thought, and so on ad infinitum.[39]

Another point to consider is whether the higher-thought account demands too much for consciousness. It allows consciousness only to creatures with the capacity for higher-order thoughts; this apparently rules out most of the animal kingdom, including human infants, from the realm of consciousness. Higher-order thoughts implicated in consciousness must have content of the form "I am in state M" or "M is occurring," where "M" refers to a type of mental state (for example, pain, the thought that your car is running low on gas). Having a thought with content "I am in state M" would require, at minimum, an ability to refer to oneself, and this in turn seems to entail the possession of some notion of self, the idea of oneself as distinct from other things and

subjects. Admittedly, all this is a complex affair: It is not clear what sorts of general conceptual, cognitive, and other sorts of psychological capacities are involved in having self-referential thoughts. Perhaps all that is required is content of the form "M is occurring." But even to have a thought with this form of content requires the possession of the concept of M. (If you have the thought that you see a tree, don't you have to have the concept of a tree—that is, to know what a tree is?) Unless you have the concepts of belief, how can you have the thought that you believe it is raining, something that is required for this belief to be a conscious belief? The possession of the concept of belief undoubtedly represents a pretty sophisticated level of cognitive and conceptual development.

It is highly plausible that some lower forms of animals, perhaps reptiles and fish, have sensations and perceptions and that the contents of their sensations and perceptions are phenomenally represented to them. (As Nagel would say, there must be something it is like for a bat to echolocate a fluttering moth.) But how plausible is it to suppose that these animals have the cognitive capacity to form thoughts of the sort required by the higher-order thought account of consciousness? It is not clear that we would want to attribute intentional states, like beliefs and thoughts, to such creatures.[40] Would we for that reason deny consciousness to such animals? It may be one thing to have *conscious* sensations and quite another to have *thoughts about* such sensations. On the face of it, the latter would seem to require a higher and more complex set of cognitive capacities than the former. The gist of the difficulty, then, is this: The higher-order thought account of consciousness makes the capacity for intentional states—of a fairly sophisticated sort—a prerequisite for having conscious states. More specifically, in order to have a conscious X, where X is a type of conscious state (like pain), the higher-order thought theory apparently requires the subject to have the concept of X. This seems like an excessive requirement.

And there are cases that seem to show that a higher-order thought does not suffice to make a mental state conscious. If after several sessions with your therapist you become aware of your hidden hostile feelings toward your roommate, that doesn't seem enough to make your hostility a conscious state. You now believe, and perhaps know, that you harbor resentments toward him, but this need not turn your resentments into conscious resentments. For this to happen, you must start to *experience*, or *feel*, these feelings, and that is not the same thing as merely thinking or believing that you have them. Another such case is the following: Suppose an injured athlete is told that she must be having a bad pain in her foot. Isn't it possible for her to believe this and yet not experience any pain? Can't she respond, "Well, maybe so, but my foot feels all right"?

Finally, we will briefly consider a possible difficulty that applies to all versions of the higher-order approach, the so-called rock objection. Alvin Goldman puts his puzzle this way:

> How could possession of a meta-state confer subjectivity or feeling on a lower-level state that did not otherwise possess it? Why would being an intentional object or referent of a meta-state [a higher-order state] confer consciousness on a first-order state? A rock does not become conscious when someone has a belief about it. Why should a first-order psychological state become conscious simply by having a belief about it?[41]

The higher-order theorist has an initial reply: The higher-order conception is intended to apply only to mental states, namely, that a mental state is a conscious state in case there is a higher-order thought or perception of it. Rocks and trees are the wrong kind of thing for the higher-order analysis.

This does not make the difficulty go away; the reply strikes one as an ad hoc move that doesn't really explain anything. One would still want to know what it is about the existence of a higher-order thought or perception that makes a first-order mental state become a conscious state. When I become aware of my pain, my pain suddenly acquires its phenomenal character, a pain quale. Don't we need a more informative account of how this happens? We can think of an electromechanical robot that perceives its environment and makes use of the information gained to guide its action. There may be good reason, even a compelling one, to attribute to such a system first-order perceptual states, and even belief states. And a robot like this may very well be equipped with an internal monitoring system that scans and monitors its first-order perceptual states. Would we for this reason say that the robot's perceptual states are conscious, or that the robot, in virtue of its self-monitoring capability, is a conscious being? We can think of various replies the higher-order theorist can try, but it seems that the point needs to be addressed.[42]

Transparency of Experience and Qualia Representationalism

Suppose you are looking at a ripe tomato, in good light. You have a visual experience with certain qualitative characters, or qualia—say, redness and roundness. Focus your attention on these qualia and try to determine the exact hue of the color, the precise shape you see, and so on—that is, try to closely inspect the qualities characterizing your experience. When you do

this, some philosophers say, you will find yourself focusing on and examining the qualities of the tomato out there in front of you. Your visual experience of the tomato is "diaphanous," or "transparent," in that when you try to introspectively examine it, you seem to look right through it to the properties of the object seen, namely, the tomato. This supposed phenomenon is called the "diaphanousness" or "transparency" of experience.[43]

Such phenomena have led some philosophers to explore an approach to qualia based on the idea that qualia are the representational contents of experiences and that the represented contents are the properties of the external objects. The view that qualia are essentially representational is called *qualia representationalism*. The red quale of your visual experience of the tomato is what your experience represents the color of the tomato as being, and when the representation is veridical, the quale is the actual red color of the tomato. So the kind of approach being described is also an *externalism* about qualia—qualia are the properties that external objects are represented as having and hence they *are* the properties of these objects when the representation is correct or accurate. This position, which locates qualia out there in the world, would enable us to reject qualia as privately introspectible qualities of inner experiences, making the approach particularly welcome to those who are committed to a physicalist stance on consciousness.

It is important to keep the following point in mind: Most everyone would accept the thesis that conscious states, at least most of them, are representational states with content. Your visual percept of a green cucumber represents the cucumber and its color, and may have, or give rise to, propositional content "There is a green cucumber there" or "I seem to see a green cucumber." According to the representationalist about qualia, however, all there is to a quale is its status as representational content, and what makes a mental state a qualitative state is its representational property. One can accept the claim that qualitative states have representational content but deny that what makes them qualitative states, states with qualia, is the fact that they represent the things they represent.

But how can the representational story about qualia be true? Aren't qualia, by definition, the qualities of your conscious experiences? How could these qualities be found in the external things around you, like tomatoes and cucumbers? Hasn't Nagel convincingly argued that you and I can have no cognitive access to what it is like to be a bat and that qualia characterizing bats' experiences are beyond our conceptual and cognitive reach? But according to the qualia externalist-representationalist, we have been misled into looking in the wrong place—we are trying to peek into a bat's mind to see what qualia are lurking in it. The idea is not just hopeless but incoherent.

So where should we look? The qualia externalist tells us to look at the external environment of the bats and try to see what objects the bats are representing and what properties the bats are representing the objects as having. Speaking of a marine parasite that attaches itself to a host only if the host's temperature is 18°C, Fred Dretske, an able proponent of the position, writes:

> If you know what it is to be 18°C, you know how the host feels to the parasite. You know what the parasite's experience is like as it "senses" the host. If knowing what it is like to be such a parasite is knowing how things seem to it, how it represents the objects it perceives, you do not have to be a parasite to know what it is like to be one. All you have to know is what temperature is. . . . To know what it is like for this parasite, one looks, not in the parasite, but at what the parasite is "looking" at—the host [to whom the parasite has attached itself].[44]

How can looking at the host the creature is "looking" at, or representing, help us find out what it is like for it to experience the temperature of 18°C? It seems that a line of consideration like the following has been influential in motivating the externalist-representationalist approach. We begin with a conception of what "qualia" are supposed to be:

> (1) Qualia are, by definition, the way things *seem*, *look,* or *appear* to a conscious creature.

So if a tomato looks red and round to me (that is, my visual experience represents the tomato as being red and round), redness and roundness are the qualia of my visual experience of the tomato. This is a representationalist interpretation of qualia. We will see how this representational view leads to qualia externalism.[45]

> (2) If things ever are the way they look or appear, qualia are exactly the properties that the perceived or represented object has. If a perceptual experience represents an object to be F (for example, the object looks F to you), and if this experience is veridical (true to the facts), then the object *is* F.

This seems reasonable: If things really are the way they are represented in perception, they must have the properties that they are perceived to have. If the tomato is the way it is represented in your visual experience, and if it is

represented as being red and round, it must really be red and round. This sounds like a tautology. What about the parasite that attaches itself to a host only when the host's temperature is 18°C? When the parasite's temperature-sensing organ is working properly and its temperature perception is veridical, the host's temperature is 18°C. The way the host's temperature is represented by the parasite is the way the temperature actually is, namely, 18°C. This means that the quale of the parasite's temperature representation is nothing other than the temperature 18°C. We have, therefore, the following conclusion:

> (3) Qualia, or phenomenal properties of experience, are among the objective properties of external objects represented in conscious experience.

To find out what it is like to be a bat in flight, zeroing in on a moth, we must shift our attention from the bat to the moth and track its fluttery trajectory through the darkness of night.

An externalist approach like this is often motivated at a deeper level by a desire to accommodate qualia within a physicalist-materialist scheme. The grapes look green to you. That is, your visual experience has the quale green. So green is instantiated. But then some *object* must instantiate it; *something* must be green. But what could this thing be? If we look inside your brain, we find nothing green there. (Even if we did find something green, how could that be what you experience?) To invoke a nonphysical mental item, like a "sense-datum" or "percept," that has the quale green goes against physicalist ontology, which tolerates no nonphysical items in the space-time world. The contents of our world are exhausted by physical-material items. Return to the question: Where is the green quale of your visual experience of the grapes instantiated? In the grapes, of course!

This answer has the virtues of boldness and simplicity, if not intuitive plausibility. Qualia representationalism-externalism has gained the support of a number of philosophers, but opinions remain sharply divided. The representationalist group considers qualia to be wholly representational—that is, qualia are fully explicable in terms of, or reducible to, their representational contents, and in line with general content externalism (chapter 8), these contents are taken to be external—that is, external to the perceiving and conscious subjects. As we said, those who disagree need not deny that qualia are representational; they can accept the view that almost all qualia, perhaps all of them, serve a representational role. What they deny is that representational contents are all there is to qualia. There are things about the qualitative features of conscious experience that are not a matter of their representing something.

What are the reasons for doubting representationalism? First, consider spectrum inversion: Spectrum inversion seems conceivable and possible. That is, where you see red, I see green, and vice versa. We both say that tomatoes are red and lettuce is green; our verbal usage coincides exactly. But the color quale of your visual experience when you look at a tomato is the same as the color quale I experience when I look at lettuce, and vice versa. We both represent tomatoes as red and lettuce as green when we assert, "Tomatoes are red and lettuce is green." But the qualia we experience are different, so representational contents cannot be all there is to qualia.[46] Second, even if within a sense modality, like vision, qualia differences and similarities amount to no more than differences and similarities in representational contents, surely there are intermodal qualia differences that are not merely differences in representational contents—for example, the qualitative differences between visual experiences and tactual experiences. We can form a belief—a representation of an external state of affairs, say, "My cat has just jumped on my lap"—on the basis of both a visual experience and a tactual experience. Isn't it obvious that there is a qualitative difference between these experiences even though the representation to which each of them leads is identical? Such examples do not silence qualia representationalists; they will try to find further representational differences between these experiences that might account for their qualitative differences. But to an anti-representationalist, piling on representational details doesn't help; it will be just more representations.

It is a natural idea to take visual experiences as representational; representing the outside world is what they do and that is their function. But there seem to be qualitative states, and qualia, that are difficult to think of as representational. Consider moods, for example: being bored, being mildly anxious or depressed, feeling upbeat, and so on. One might say that these moods "represent" something about our states of psychological well-being, or some such thing. But this is not the sense of representation at issue. Representation in the sense relevant to representationalism applies only where talk of representational *accuracy, fidelity*, and *correctness* makes sense; as was noted earlier, representations in this context must have "satisfaction conditions" and be evaluable in terms of how closely they meet these conditions. Evaluating moods in terms such as "accuracy" and "fidelity" doesn't make any sense. And how about the qualia, or what it's like aspects, that accompany emotions, like anger, embarrassment, and jealousy? Do they represent anything? Can they be "true" or "accurate"? True, or accurate, *to what*? Further, the transparency, or diaphanousness of experience, though it is not implausible for visual experience, makes little sense for moods, or the qualia involved in emotions: What is a sense of boredom transparent to?

The reader should also think about transparency of experience in connection with perceptual experiences other than visual ones—for example, whether auditory, olfactory, and tactile experiences are transparent in a similar sense.

At this point, an alternative view of the situation suggests itself. Representation requires a representational vehicle, the thing that does the representing (for example, sentences, pictures, maps), which is distinct from the object it represents (states of affairs, people and objects, the layout of cities). Qualia, or most of them, do represent, and that could only mean that they are the representational vehicles, the internal states that do the representing. As such, they are distinct from the things they represent, external objects and their properties, like tomatoes and their colors and shapes. So qualia should not be identified with the objects and properties represented by experience. Does such a view lead to an anti-physicalist view of qualia? Not necessarily. The view sits well with the psychoneural identity theory, which identifies states with qualia with neural-physical states of the brain; this theory would identify qualia with neural-physical properties of brain states. The proponents of the identity theory would say that these brain states, in virtue of their neural-physical properties, represent the external objects and their properties. Qualia, as properties of states of the brain, play a role in determining what these brain states represent; but they stay "inside."

What then of the series of considerations—(1), (2), and (3)—that appears to lead to qualia externalism? Philosophers have distinguished between two senses of "appear," "seem," and "look"—the "epistemic" (or "doxastic") sense and the "phenomenal" sense. These expressions are used in the epistemic sense when we say things like "It appears that (seems that, looks like) the Democrats will control Washington politics for years to come," and "Prospects for a compromise on the bailout bill seem (appear, look) quite bleak." In this usage, "appear," "seem," and "look" have roughly the sense of "there is reason to believe," "evidence indicates," or "I am inclined to believe." We are already familiar with the phenomenal sense of these expressions: When a tomato appears or looks red to you, your visual experience is characterized by a certain qualitative property, the quale red. Red refers to the way the tomato visually appears or looks. You are looking at a red tomato bathed in brilliant green light, and you report "That tomato looks green" even if you know that the tomato is red. So "looks green" in this context reports a visual quale, not your inclination to believe that it is green.

With this distinction in mind, let us return to (1), (2), and (3). A plausible case can be made for the observation that "appear" and "seem" are used equivocally in (1) and (2). More specifically, (1) is acceptable as a definition of qualia

only if "appear" and "seem" are used in their phenomenal, or sensuous, sense. Qualia are the ways in which objects and events around me, and in me, present themselves in my experience; they are how the yellow of Van Gogh's sunflowers appears to me, how the pain in my knee feels (it hurts), how a breeze over a lavender field in bloom smells. Now consider (2) again: "If things ever are the way they look or appear, qualia are exactly the properties that the perceived or represented object has." This statement is plausible only if "look" and "appear" in the antecedent are understood in an epistemic, or doxastic, sense—that is, to mean something like "If things are the way our perceptual experience indicates them to be." And under this supposition, it would follow that things do quite often have the properties we believe them to have on the basis of perceptual experience. So (2) requires the epistemic-doxastic sense of "appear" and "look." Now return to (1)—the supposed definition of qualia. If we read (1) with this sense of "appear" and "seem" in mind, it says something like "Qualia are the properties that we have reason to believe things to have on the basis of perceptual experience." The trouble obviously is that if (1) is interpreted this way, there is no reason to accept it as true of qualia, much less a definition of what qualia are. It is a reasonable definition of qualia only if "appear," "seem," and "look" are read in their phenomenal sense. If this is right, the chain of reasoning represented by (1), (2), and (3) is fallacious as it equivocates on "appear," "look," and "seem."

Here we have only skimmed the continuing debates between qualia representationalists and those skeptical about this approach.[47] The division between the two camps is deep and well entrenched. The debates have been intense and show no sign of abating. This is one of the most important current issues about consciousness.

For Further Reading

Torin Alter and Robert Howell's *A Dialogue on Consciousness* is a short and accessible discussion that touches on most of the important issues on consciousness, in an entertaining dialogue form. Robert Van Gulick's "Consciousness" in the *Stanford Encyclopedia of Philosophy* is a useful resource. *The Blackwell Companion to Consciousness*, edited by Max Velmans and Susan Schneider, is an up-to-date collection of essays by philosophers and scientists on a wide range of philosophical and scientific issues on consciousness.

Janet Levin's "Could Love Be Like a Heatwave?" is an interesting and helpful discussion of Nagel's paper on bats (Levin's paper also takes up Frank Jackson's "knowledge argument," which is covered in chapter 10).

For higher-order theories, Peter Carruthers's "Higher-Order Theories of Consciousness" in the *Stanford Encyclopedia* is a clear and balanced survey and discussion; you may also consult his *Consciousness: Essays from a Higher-Order Perspective*. Leopold Stubenberg's *Consciousness and Qualia* has an interesting and informative chapter on higher-order theories (chapter 4). *Consciousness and Mind* by David Rosenthal includes important papers on the higher-order thought theories.

For representationalist accounts of consciousness and qualia, see Fred Dretske, *Naturalizing the Mind*; Christopher Hill, *Consciousness*; Michael Tye, *Ten Problems of Consciousness*; and William Lycan, *Consciousness and Experience*. Also interesting is Alex Byrne's "Intentionalism Defended." Hill's recent book presents a sophisticated formulation and defense of the representational approach. For critical discussion, see Ned Block, "Mental Paint." In particular on the alleged transparency of perceptual experience, see Amy Kind, "What's So Transparent about Transparency?" and "Restrictions on Representationalism"; Charles Siewert, "Is Experience Transparent?"

Ned Block's *Consciousness, Function, and Representation* collects many of his influential papers on consciousness.

The Nature of Consciousness: Philosophical Debates, edited by Ned Block, Owen Flanagan, and Güven Güzeldere, is a comprehensive and indispensable anthology of articles on consciousness; the topics discussed in this and the following chapter are well represented.

Notes

1. René Descartes, *Meditations on First Philosophy*, Meditation VI.

2. René Descartes, "Author's Replies to the Fourth Set of Objections," p. 171.

3. Ivan Pavlov, *Experimental Psychology and Other Essays*. It is clear that by "psychic life" Pavlov had in mind conscious life.

4. Daniel C. Dennett, "Quining Qualia." Emphasis added.

5. Reported in Dennett's *Consciousness Explained*, p. 383.

6. Georges Rey, "A Question about Consciousness."

7. T. H. Huxley, *Lessons in Elementary Physiology*, p. 202.

8. William James, *The Principles of Psychology*, p. 647 in the 1981 edition.

9. For example, do a search on Internet book sites, such as Amazon and Barnes & Noble, with the keywords "consciousness" and "mystery."

10. Thomas Nagel, "What Is It Like to Be a Bat?" p. 528 in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

11. Francis Crick, *The Astonishing Hypothesis*, p. 19.

12. Patricia S. Churchland, "Can Neurobiology Teach Us Anything about Consciousness?" in *The Nature of Consciousness*, ed. Block, Flanagan, and Gülzedere, p. 138.

13. It is interesting to note that of the forty-nine chapters collected in the eight hundred–page anthology of central modern texts on consciousness (*The Nature of Consciousness* [1999], ed. Block, Flanagan, and Güzeldere), only one chapter predates Nagel's "Bat" paper. It is "The Stream of Consciousness" by William James, published in 1910. Although Nagel's paper's impact on the science of consciousness is less clear, it is a fact that consciousness began making a big comeback in science and philosophy at about the same time.

14. This is a bit of an overstatement. It seems that we can know, and Nagel would agree, that what it is like to be a flying bat echolocating a moth isn't similar to what it is like to be, say, a frightened rabbit or a stampeding elephant, and lots of other similar things.

15. Over the years, Nagel has been much interested in the subjective-objective contrast, which he explains in terms of "point of view." It is fair to say that the idea of a point of view is the central concept that shapes the arguments throughout Nagel's book *The View from Nowhere*.

16. Simulation theory about folk-psychological attributions of mental states holds that such empathetic "mind reading" does take place among people; in fact, there may be a biological basis for this phenomenon. See Alvin Goldman, *Simulating Minds*. Note, however, that simulation theory focuses on intentional states, like beliefs, goals, plans, and decisions, while our current interest concerns the qualitative character of conscious experience.

17. This should not be taken to imply that these are the only kinds of consciousness. Christopher Hill distinguishes five "forms" of consciousness, in chapter 1 of his *Consciousness*. But phenomenal and access consciousness, roughly in the sense to be explained, appear in Hill's typology of consciousness.

18. "Qualia" is now the standard term. Sometimes it is used to refer to states with such qualitative characters.

19. Ned Block, "On a Confusion about a Function of Consciousness," in *The Nature of Consciousness*, ed. Block, Flanagan, and Güzeldere, p. 377.

20. Christopher Hill, *Consciousness*, p. 21.

21. It has been claimed, plausibly, that very often when you are asked "Do you believe that *p*?" you don't look into your mind and try to determine if you are in a certain mental state; rather, what you do is to try to see if *p* is true. Think about being asked at an airport lounge "Do you believe our flight is on time?"

22. If anger, say, is *partially constituted* by a quale ("anger quale"), then all instances of anger must exhibit this anger quale; this quale is part of what makes it an instance of anger. But it may well be that though each instance of anger has a certain quale, there is no single anger quale present in all instances of anger.

23. Ned Block, "On a Confusion about a Function of Consciousness," in *The Nature of Consciousness*, ed. Block, Flanagan, and Güzeldere, p. 382.

24. Bernard Baars, *In the Theater of Consciousness.*

25. Daniel Dennett, *Consciousness Explained.*

26. Ned Block, "On a Confusion about a Function of Consciousness," in *The Nature of Consciousness*, ed. Block, Flanagan, and Gülzedere, p. 385. Another possible example is philosophical "zombies." Though zombies, by definition, lack phenomenal consciousness, they can have access-conscious states, states with representational contents that guide their behavior. Whether zombies are metaphysically possible is a disputed question. More on zombies in chapter 10.

27. Lawrence Weiskrantz, *Blindsight*.

28. See also John R. Searle, *The Rediscovery of the Mind*, especially chapter 4.

29. This was discussed in some detail in chapter 1.

30. See Richard Nisbett and Timothy DeCamp Wilson, "Telling More Than What We Can Know." Also Alison Gopnik. "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality."

31. Thomas Nagel, "What Is It Like to Be a Bat?" p. 437. Emphasis added.

32. Thomas Nagel, "Subjective and Objective," p. 201. Emphasis added.

33. Nagel's point may seem to go further, in its emphasis on there being a "single" subject for each experience—that is, at least one and *at most one* subject. Why can't an experience belong to two or more subjects? If we take experiences to be states of a subject, there may be a simple answer: If X and Y are distinct things, X's states must be distinct from Y's states; that is, X's being in a certain state must be distinct from Y's being in some state since X and Y are constituents of their respective states.

34. See Christopher Hill, *Consciousness*, p. 19.

35. Earlier we raised doubts about a possible belief quale on the ground that no phenomenal quale could exist for unconscious mental states. The present paragraph might be taken to imply that a pain of which a subject is unaware could exist—that is, there can be unconscious pains. Does this contradict the earlier argument? Not necessarily. Note that the claim is not that unconscious phenomenal pains could exist. The matter may be a little complicated, though; nothing important hinges on it, and the reader may simply bracket it.

36. David Armstrong, "The Nature of Mind," p. 198.

37. David Rosenthal, "The Independence of Consciousness and Sensory Quality," p. 31. See also Rosenthal, "Explaining Consciousness."

38. Gilbert Harman, "The Intrinsic Quality of Experience."

39. Alvin I. Goldman, "Consciousness, Folk Psychology, and Cognitive Science"; Mark Rowlands, "Consciousness and Higher-Order Thoughts."

40. For an argument, see Donald Davidson, "Rational Animals."

41. Alvin I. Goldman, "Consciousness, Folk Psychology, and Cognitive Science," in *The Nature of Consciousness*, ed. Block, Flanagan, and Güzeldere, pp. 112–113.

42. In preparing this section, I am indebted to Peter Carruthers's "Higher-Order Theories of Consciousness."

43. See Gilbert Harman, "The Intrinsic Quality of Experience," and Michael Tye, *Ten Problems of Consciousness*, pp. 30–31 (the "problem of transparency"). Hill's *Consciousness* contains informative and helpful discussions of transparency; see chapters 2 and 3. Also see "For Further Reading."

44. Fred Dretske, *Naturalizing the Mind*, p. 83.

45. Can there be an internalist form of representationalism? The answer is not clear. For discussion, see Ned Block, "Mental Paint."

46. If you think interpersonal spectrum inversion makes no sense, we can imagine spectrum inversion in the same person over time. One morning you wake up and realize that things around you look different in color than you remember from the day before—tomatoes look green and spinach looks red. And your friends assure you that nothing has changed color from the day before; things look just the same to them. What are you to think? See Sydney Shoemaker, "Inverted Spectrum." See also Martine Nida-Rümelin, "Pseudo-Normal Vision: An Actual Case of Qualia Inversion?"

47. In the next chapter we will briefly discuss Hill's representational theory of pain, the "bodily disturbance" theory.

CHAPTER 10

Consciousness and the Mind-Body Problem

The central focus of this book has been the mind-body problem, the problem of clarifying and understanding how our minds are related to, or grounded in, our bodily nature. It is fair to say, though, that this problem ultimately comes down to understanding how our conscious life is related to the biological-physical processes going on in our brain. That is, the core of the mind-body problem is the consciousness-brain problem. There seems little question that, so far as we know, conscious states depend on, or arise from, the physicochemical processes in the brain. But how do the electrochemical processes in the gray matter of the brain give rise to an awareness of colors, shapes, motions, sounds, smells, and other sensory qualities, delivering to us a rich, kaleidoscopic picture of the world around us? As many have observed, consciousness is what makes the mind-body problem so hard—perhaps "hopeless," as Thomas Nagel has said.

Materialism, or its contemporary successor, physicalism, is the default position in modern science and much of contemporary philosophy of mind (at least, in the analytic tradition). The world we live in is an essentially physical world; physical processes seem to underlie all events and processes, and it isn't for nothing that we think of physics as our basic science, a science that aspires to a "full coverage"[1] of the world. Can minds and consciousness be accommodated in such a world? If physicalism is true and consciousness is real, there must be such an accommodation; consciousness must have a well-defined place in the physical world. But is that possible? There is a general consensus that the fate of physicalism hangs on whether consciousness can be given an adequate physical account. Physical science has been able to explain the phenomenon of life—from molecular genetics we now know how reproduction, arguably the most salient

biological phenomenon, is possible. One of the "two great mysteries" of the world, life and mind, has yielded to physical explanation. Can we expect the same for the one remaining mystery, mind and consciousness?

In this chapter, we examine a cluster of issues concerning consciousness, the mind-body problem, and physicalism.

The "Explanatory Gap" and the "Hard Problem"

As we saw earlier (chapter 1), most philosophers accept mind-body supervenience in something like the following form:

> If an organism is in some mental state M at *t*, there must be a neural-physical state P such that the organism is in P at *t*, and any organism that is in P at any time is necessarily in mental state M at the same time.

Briefly, the thesis is that every mental state has an underlying "supervenience base" in neural states. Consider pain. According to mind-body supervenience, whenever you are in pain, there is a neural state that is the supervenience base of your pain. Call this neural state N. Whenever N occurs, you experience pain, and we may suppose that unless N occurs, you do not experience pain. We also say things like: Pain "arises out" of N, or "emerges" from N, or N is a "neural substrate" or "correlate" of pain.[2]

But why is it that pain, not itch or tickle, occurs when neural state N occurs? What is it about the neural-biological-physical properties of N that make it the case that pain, not another kind of sensory experience, arises when N occurs? Further, why does pain not arise from a different neural state? Why does any conscious experience arise from N? Here we are asking for an explanation of why the pain-N supervenience relation holds. The problem of the explanatory gap is that of providing such an explanation—that is, the problem of closing the apparent "gap" between pain and N or, more generally, between phenomenal consciousness and the brain.[3]

If there is indeed an explanatory gap here, its existence does not depend on using the idiom of supervenience. Even if you want to limit yourself to talk of psychophysical correlations, the problem still arises. Suppose pain correlates with neural state N. Why does pain correlate with N rather than another neural state? Why doesn't itch or tickle correlate with N? If pain correlates with N and itch with a different neural state N*, there must be an explanation of this

fact, we feel, in terms of the neural-physical differences between N and N*. What would such an explanation look like? How would anyone go about finding one? Could neurobiological research ever discover explanations of these and other psychoneural correlations? If we don't have such explanations yet, what *further* scientific research would help us meet our explanatory needs? More generally, the question is this: Why do conscious states correlate with the neural states with which they correlate?[4] Or, why do conscious states supervene on the neural states on which they supervene?

Although the term "explanatory gap" is relatively new, the problem is not. As we saw in the preceding chapter, William James wrote well over one hundred years ago:

> According to the assumptions of this book, thoughts accompany the brain's workings, and those thoughts are cognitive of realities. The whole relation is one which we can only write down empirically, confessing that no glimmer of explanation of it is yet in sight. That brains should give rise to a knowing consciousness at all, this is the one mystery which returns, no matter of what sort the consciousness and of what sort the knowledge may be. Sensations, aware of mere qualities, involve the mystery as much as thoughts, aware of complex systems, involve it.[5]

James recognizes that thoughts and sensations correlate with ("accompany") brain processes, but we can only make a list of these correlations ("write down empirically," as he says). Keeping a running list of observed psychoneural correlations hardly amounts to understanding why these correlations hold—or why there are psychoneural correlations at all. The list seems brute and arbitrary; there seems no reason why these particular correlations, not another arbitrarily permuted set of them, should hold. According to James, it is one "mystery" for which there is no "glimmer of explanation." You will recall another noted scientist, T. H. Huxley, expressing a similar sentiment even earlier, saying, "How it is that anything so remarkable as a state of consciousness comes about as the result of irritating nervous tissue is just as unaccountable as any other ultimate fact of Nature."[6]

More recently, the problem of phenomenal consciousness has also been called the "hard problem" of consciousness. According to substance physicalism, a position that rejects immaterial minds as bearers of mental properties, it is physical systems, like biological organisms, that have mental properties—that is, have beliefs and desires, learn and remember, experience pains and remorse, are upset and fearful, and all the rest. Now consider this question

posed by David Chalmers: "How could a physical system be the sort of thing that could *learn*, or that could *remember*?"[7] Chalmers calls this an "easy" problem—a tractable component of the mind-body problem. As he acknowledges, the scientific problem of uncovering the details of the neural mechanisms of memory may present the brain scientist with formidably difficult challenges. And yet there is here a well-defined research project: Identify the underlying neural mechanisms—say, in humans and higher mammals—that process information received from perceptual systems, store it, and retrieve it as needed. The problem, although not easy from a scientific point of view, does not seem to pose any special puzzles or mysteries from the philosophical point of view; conceptually and philosophically, it is an "easy" problem.

The reason for this is that memory is a "functional" concept, a concept defined in terms of the job that memory performs in the cognitive-psychological economy of an organism (see chapter 5). So the question "How could a physical system manage to remember?" seems to have answers of the following straightforward form: A physical system with neural mechanism N can remember, because to remember is to perform a set of tasks T, and neural mechanism N enables a system to perform T; moreover, we can explain exactly how N performs T in this system and others like it. Identifying mechanism N, for a population (humans, mammals) under investigation, is a scientific research project, and the functional characterization of remembering in terms of T is what makes it possible to define the research program. There is no special philosophical mystery here, or so it seems.

Compare this situation with one that involves qualitative states of consciousness. That is, instead of asking, "How could a physical system be the sort of thing that could *remember*?" ask, "How could a physical system be the sort of thing that could *experience pain*?" This is what Chalmers calls the "hard problem"—the hard, and possibly intractable, part of the mind-body problem. The problem is hard because pain apparently resists a functional characterization. Granted, pains have typical input conditions (tissue damage and trauma) and typical behavioral outputs (winces, groans, avoidance behavior). But many of us are inclined to believe that what makes pain pain is that it is experienced as painful—that is, pain hurts. This phenomenal, qualitative aspect of pain seems not capturable in terms of any particular task associated with pain. Uncovering the neural mechanism of pain—that is, the neural mechanism that responds to tissue damage and triggers characteristic pain responses—is the "easy" part, although it probably is a challenging research problem for the brain scientist. What is "hard" is the problem of an-

swering further questions like "Why is pain experienced when this neural mechanism is activated? What is it about this mechanism that explains why pain rather than itch, or anything at all, is experienced when it is activated?"

The "hardness" of the hard problem can be glimpsed from the following fact: The question "How can neural mechanism N enable a system to perform task T?" where T is the cluster of tasks associated with remembering, seems answerable *within* neurophysiology and associated physical-behavioral sciences. In contrast, "How can neural mechanism N enable a system to experience pain?" does not seem answerable within neurophysiology and associated sciences, because "pain," or the concept of pain, does not even occur in neurophysiology or other physical-behavioral sciences. Pain, as the term is used here, is a phenomenally conscious event; what it is like to experience pain as opposed to, say, what it is like to experience an itch or see yellow, is of the essence of pain in this sense. Pain as a type of phenomenal consciousness, therefore, lies outside the scope of brain science. Given this, it is difficult to see how there could be a solution to our problem within brain science. If neuroscience does not have the expressive resources even to talk about pain (as a quale), how can it explain why pain correlates with a neural state with which it correlates?

That the explanatory gap cannot be closed, or that the hard problem cannot be solved, is the central theme of emergentism. Psychoneural correlations are among the ultimate unexplainable brute facts, and Samuel Alexander and C. Lloyd Morgan, leading emergentists of the early twentieth century, counseled us to accept them with "natural piety"—stop asking why and just be grateful that consciousness has emerged!

But should we give up the hope for explanations of psychoneural correlations for good and accept the explanatory gap as unclosable and the hard problem as unsolvable? What exactly would it take to give a physical-neural account of phenomenal consciousness? And what would it take to deal with the explanatory gap? These questions are now often posed in terms of *reduction* and *reductive explanation*. The thought is that to achieve a solution to the hard problem and close the explanatory gap, we must be able to *reduce* consciousness to neural states, or *reductively explain* consciousness in terms of neural processes. We will see below what options there are for reducing or reductively explaining phenomenal consciousness.

People have spoken of the miracle, and mystery, of the mind. The mystery of the mind is, in essence, the mystery of consciousness. And consciousness may well be what makes the mind truly miraculous. Other aspects of mentality, like belief, emotion, and action, may have explanations along the line of

the functional-neural account of memory described earlier. However, phenomenal experiences, or qualia, apparently present us with an entirely different problem not easily amenable to an account of a similar sort.

Does Consciousness Supervene on Physical Properties?

Self-awareness, in the sense of awareness of our own psychological states, seems in principle explicable in terms of some internal monitoring mechanism or higher-order perception or thought (see chapter 9), and this provides us with a basis for a possible physical-neural explanation of self-awareness. The "directness" and "immediacy" of such awareness perhaps can be explained in terms of a direct coupling of such scanning devices to other cognitive modules and to the speech center, a mechanism responsible for verbal reports. The first-and third-person asymmetry of access may be no great mystery: It arises from the simple fact that my scanning device (and its associated speech center), not yours, is directly monitoring my internal states.[8] These ideas are rough and may ultimately fail. However, what they show is the possibility of understanding the subjectivity of consciousness in the sense of direct first-person access to one's own mental states, because at least we can imagine a possible mechanism that can implement it at the physical-biological level. We can see what it would be like to have an explanation of certain important aspects of the subjectivity of consciousness. This shows that at least we understand the problem.

On this view of consciousness as direct awareness of internal states, consciousness would be supervenient on the basic physical-biological structure and functioning of the organism. The fact that an organism is equipped with a capacity for direct monitoring of its current internal states is a fact about its physical-biological organization and must be manifested through the patterns of its behavior in response to input conditions. This means that if two organisms are identical in their physical-biological makeup, they cannot differ in their capacity for self-monitoring. In that sense, consciousness as special first-person epistemic authority may well be supervenient on physical and biological facts.

What, then, of the phenomenal, qualitative aspect of consciousness? Do qualia supervene on the physical-biological constitution of organisms? You feel pain when your C-fibers are stimulated; is it necessarily the case that your physical duplicate feels pain when her C-fibers are stimulated? In our world,

pains and other qualitative states exist, and we suppose them to depend, in regular lawlike ways, on what goes on in the physical-biological domain. Is there a possible world that is a total physical duplicate of this world but in which there are no phenomenal mental states? Some influential philosophers think that such worlds are possible. For example, Saul Kripke writes:

> What about the case of the stimulation of C-fibers? To create this phenomenon, it would seem that God need only create beings with C-fibers capable of the appropriate type of physical stimulation; whether the beings are conscious or not is irrelevant here. It would seem, though, that to make the C-fiber stimulation correspond to pain, or be felt as pain, God must do something in addition to the mere creation of the C-fiber stimulation; He must let the creatures feel the C-fibers as pain, and not as a tickle, or as warmth, or as nothing, as apparently would also have been within His powers.[9]

Kripke contrasts this situation with one involving molecular motion and heat: After God created molecular motion, he did not have to perform an additional act to create heat. When molecular motion came into being, heat came along with it.

If Kripke is right, there are two kinds of possible worlds that, though identical with our world in all physical respects, are different in mental respects: First, there are worlds with different physical-phenomenal correlations (for example, C-fiber stimulation correlates with itches rather than pains), and second, there are those in which there are no phenomenal mental events at all—"zombie worlds." In the latter, there are creatures exactly like you and me in all physical respects, behaving just as we do (including making noises like "That toothache kept me awake all night, and I am too tired to work on the paper right now"), but they are zombies with no experience of pain, fatigue, sensing green, or any of the rest. Their inner lives are dark and empty Cartesian theaters.

But is Kripke right? How is it possible for God to create C-fiber stimulation but not pain? Let us look at various considerations against qualia supervenience:

1. There is no conceptual connection between the concept of pain and that of C-fiber stimulation (that is, no connection of meaning between the terms "pain" and "C-fiber stimulation"), and therefore there is at least no logical contradiction in the supposition that an organism has its C-fibers stimulated without experiencing pain or any

other sensation. This argument will be challenged on the following ground: There is no conceptual connection between heat and molecular motion either, nor between water and H_2O, and yet there is no possible world in which molecular motion exists but heat does not, or a world that contains H_2O but no water. This shows that the absence of logical or conceptual connection does not prove it is possible to have one without the other. In considering this reply, we must ask whether the case of pain and C-fiber excitation is relevantly similar to cases like heat-molecular motion and water-H_2O.

2. "Inverted spectra" are possible: It seems perfectly conceivable that there are worlds that are just like ours in all physical respects but in which people, when looking at the things we look at, experience colors that are complementary to the colors we experience.[10] In such worlds, cabbages are green and tomatoes are red, just as they are in this world; however, people there experience red when they look at cabbages and green when they look at tomatoes, although they call cabbages "green" and tomatoes "red." Such worlds seem readily conceivable, with no hidden contradictions. In fact, why isn't it conceivable that there could be a world in which colors are sensed the way sounds are sensed by us and vice versa—worlds with inverted sense modalities? (Arthur Rimbaud, the French poet, saw colors in vowels, like this: "A black, E white, I red, U green, and O blue.")[11]
3. In fact, why couldn't there be people in our world, perhaps among our friends and family, whose color spectra are inverted with respect to ours, although their relevant neural states are the same? They call tomatoes "red" and cabbages "green," as we do, and all our observable behaviors coincide perfectly. However, their color experiences are different from ours.[12] We normally do not imagine such possibilities; we assume that when you and I are in relevantly similar perceptual situations, we experience the same qualitative sensation. But this is precisely the assumption that sensory states are determined by physical conditions, and that is what is at issue. What matters to our shared knowledge of the world and our ability to coordinate our actions is that we can discriminate the same range of colors, not how these colors appear to us (this point is further discussed below).
4. Implicit in these remarks is the point that qualia also do not supervene on functional properties of organisms.[13] A functional property is, roughly, a property indicating how an organism responds to a given

> sensory input by emitting some specific behavior output. You and your physical duplicate must share the same functional properties—that is, functional properties supervene on physical properties (think about why this must be so). So if qualia should supervene on functional properties, they would supervene on physical properties. Therefore, to question the supervenience of qualia on physical properties is ipso facto to question their supervenience on functional properties.[14]

The main argument for the failure of the physical supervenience of qualia, then, is the apparent conceivability of zombies and qualia inversions in organisms physically indistinguishable from us.[15] Conceivability may not in itself imply real possibility, and the exact relationship between conceivability and possibility is a difficult and contentious issue. Moreover, we could make errors in judging what is conceivable and what is not, and our judgments may depend on available empirical information. Knowing what we now know, we may not be able to conceive a world in which water is not H_2O, but people who did not have the same information might have judged differently. In the case of qualitative characters of mental states, however, is there anything about them that, should we come to know it, would convince us that zombies and qualia inversions are not really possible? Don't we already know all we need to know, or can know, about these subjective phenomenal characters of our experience? Is there anything about them that we can learn from objective, empirical science? Research in neuropsychology will perhaps tell us more about the biological basis of phenomenal experiences, but it is difficult to see how that could be relevant as evidence for qualia supervenience. Such discoveries will tell us more about correlations between qualia and underlying neural states, but the question has to do with whether these correlations are *metaphysically necessary*—whether there are possible worlds in which the correlations fail. What is more, it is not so obvious that neurophysiological research could even establish correlations between qualia and neural states. For the claimed correlations, we might point out, only correlate neural states with *verbal reports* of qualia, not with qualia themselves, from which it follows that these correlations may be consistent with qualia inversions (see below on consciousness and science). At best, one might argue, scientific research could only correlate neural states with the color similarities and differences a subject can discriminate, not with the intrinsic qualities of his color experiences.

It seems prudent to conclude at this point that, though the case against qualia supervenience is not conclusive, it is not insubstantial either. Are there,

then, considerations in favor of qualia supervenience? It would seem that the only positive considerations are of a broad metaphysical sort that might be accused of begging the question. Say you are already committed to physicalism: You then have two choices about qualia—either to deny their existence, as qualia eliminativists, or nihilists, would urge,[16] or to try to accommodate them somehow within a physicalist framework. Given the choice between accommodation and rejection, you opt for accommodation, since a flat denial of qualia, you feel, makes your physicalism fly in the face of common sense. You may then find supervenience an appealing way of bringing qualia into the physical fold. For qualia supervenience at least guarantees that once all the physical details of an organism are fixed, that completely fixes all the facts about its qualia. This may not make qualia full-fledged physical items, but it at least makes them fully dependent on physical facts, and that protects the primacy and priority of the physical. That, you may feel, is good enough.

Further, qualia supervenience may seem to open a way of accounting for the causal relevance of qualia. If qualia are genuine existents, their existence must make a causal difference. But any reasonable version of physicalism must consider the physical world to be causally closed (see chapter 7), and it would seem that if qualia are to be brought into the causal structure of the world, they must at least be supervenient on physical facts of the world. Supervenience by itself may not be enough to confer *causal efficacy* on qualia, but it may suffice to make them *causally relevant* in some broad sense. In any case, without qualia supervenience, there may well be no hope of providing qualia with a place in the network of causal relations of this world.[17] But this is not an argument for qualia supervenience; it means only that qualia supervenience is on our wish list. For all we know, qualia might be epiphenomenal, without powers to cause anything. This cannot be ruled out a priori, and it would not be proper to use its denial as a premise in a philosophical argument.

The question of qualia supervenience presents a deep dilemma for physicalism. If qualia supervene on physical-biological processes, why they supervene—why they arise from the specific neural substrates from which they in fact arise—remains a mystery, something that seems inexplicable from the physical point of view. Yet if qualia do not supervene, they must be taken as properties outside the physical domain, not answerable to physical laws. At this point, some may feel that the basic physicalist approach to mentality is in deep trouble and that it is time to begin exploring nonphysicalist alternatives. But are there real alternatives to physicalism? For most contemporary philosophers of mind, Cartesian substance dualism is not a live option. In fact, it is difficult to see what

concrete help could be expected from substance dualism about the problems of consciousness and other outstanding issues about the mind. If our goal is to build a picture of the world in which the conscious mind occupies a natural and intelligible place, then without qualia supervenience we seem faced with a wide-open landscape with little guidance as to the direction we ought to take.

Closing the Explanatory Gap: Reduction and Reductive Explanation

How might the explanatory gap be closed? How might the hard problem of consciousness be solved? We will consider here two ways of attempting answers to these questions, reduction and reductive explanation. The idea is that if we can either reduce, say, pain to C-fiber stimulation, by identifying them (as in psychoneural identity theory), or reductively explain pain phenomena in terms of C-fiber stimulations and laws at the neural level, that should suffice to close the explanatory gap and resolve the hard problem.

How might we reductively explain consciousness in terms of neural processes? Someone who favors this approach to the explanatory gap is apt to do so because of the thought that reductive explanation is possible even where reduction is not. David Chalmers writes:

> In a certain sense, phenomena that can be realized in many different physical substrates—learning, for example—might not be reducible in that we cannot *identify* learning with any specific lower-level phenomena. But this multiple realizability does not stand in the way of reductively explaining any instance of learning in terms of lower-level phenomena.[18]

Chalmers apparently takes the multiple realizability of learning to rule out the identification of learning with some particular neural-biological process—that is, an *identity reduction* of learning—but in his view this does not preclude a reductive explanation of the phenomenon.

Jerry Fodor, a tireless critic of reductionism, appears to be driving at the same point when he writes:

> The point of reduction is *not* primarily to find some natural kind predicate of physics coextensive with each kind predicate of a special science. It is, rather, to explicate the physical mechanisms whereby events conform to the laws of the special sciences.[19]

In his first sentence, Fodor is saying that the point of reduction is not to find the so-called bridge laws connecting special-science predicates with physical predicates—laws of the kind that used to be thought to be required for reduction. In Fodor's view, the phenomenon of multiple realization makes such laws unavailable, but that doesn't really matter. His positive view is that genuine reduction consists in producing reductive explanations of higher-level phenomena in terms of underlying "physical mechanisms," and that such reductive explanations do not require bridge laws.

This raises some interesting questions. What is a reductive explanation, and how does it work? How does a reductive explanation differ from an ordinary, nonreductive explanation? How are reduction and reductive explanation related to each other? Is reductive explanation really possible without reduction? Does reduction always give us reductive explanation?

One point we can be certain of is that mere psychophysical correlation laws or bridge laws, or mind-body supervenience relations, do not generate reductive explanations—an understanding of mental phenomena in terms of their underlying neural-biological phenomena. Suppose that we have in hand the correlation law connecting pain and C-fiber stimulation (Cfs), and consider the following derivation as a possible explanation:

(α) Jones is in Cfs state at *t*.
A person is in pain at *t* iff she is in Cfs state at *t*.
Therefore, Jones is in pain at *t*.

The second line of this derivation is the correlation law connecting pain with Cfs. It allows us to derive a fact about Jones's consciousness from a fact about her brain state. We can grant that (α) is an explanation of some sort: It has the form of so-called deductive-nomological explanation—that is, an explanation in which a statement of the event to be explained is derived from antecedent conditions together with laws.[20] But is it a *reductive* explanation, one that gives us an understanding of how or why pain arises from neural processes, thereby helping to close the explanatory gap?

The answer: Definitely not. The problem lies with the pain-Cfs correlation law used as a premise of the derivation. When we want a reductive understanding of conscious states on the basis of neural processes, we want to know how sensations like pain and feelings like distress arise out of neural states—or why these conscious states correlate with the neural states with which they correlate. Why does pain correlate with Cfs, not another neural state? Why doesn't another phenomenal state, say tickle, correlate with Cfs? Instead of at-

tempting to answer these questions, (α) simply assumes the pain-Cfs correlation law as an unexplained premise of the explanation. To put it another way, psychoneural correlation laws are exactly what need to be explained if we are to gain a reductive understanding of consciousness in brain science. You may recall William James's despairing of ever gaining an explanatory insight into why thoughts and sensations "accompany" the neural states that they do. In speaking of sensations "accompanying" states of the brain, James is acknowledging that there are psychoneural correlation laws and that we know at least some of them. That is not the issue. According to James, what we need but do not have as yet (and perhaps never will) is an explanation of these correlations. Only such an explanation would dispel the "mystery" of consciousness and close the explanatory gap.

What if these psychoneural correlation laws could be strengthened into psychoneural identities? What if instead of "pain occurs iff Cfs occurs," we had "pain = Cfs"? That is, suppose we had *identity reductions* of conscious states to neural states. How might that help us close the explanatory gap? Would it generate a reductive explanation of pain in terms of Cfs? Consider then the following derivation:

(β) Jones is in Cfs state.
Pain = Cfs.
Therefore, Jones is in pain.

There is no question that (β) is a valid argument: The conclusion is obtained by putting "equals for equals" in the first premise. But is it any sort of explanation? The answer has to be in the negative. The best way of understanding what is going on in this derivation is to see that the conclusion is nothing but a "rewrite" of the premise, with the identity "pain = Cfs" sanctioning the rewriting. Given that pain = Cfs, the fact stated by "Jones is in Cfs" is the very same fact that is stated by "Jones is in pain"; the conclusion states no new fact over and above what is stated in the premise.

Derivation (β) is no more explanatory than a derivation like this:

(β*) Tully rebuked Catiline.
Tully = Cicero.
Therefore, Cicero rebuked Catiline.

No one would take this seriously as an explanation of why Cicero rebuked Catiline. If these remarks are on the right track, how does identity reduction

help with the explanatory gap and the hard problem? How does the identity "pain = Cfs" help us deal with the question "Why does pain correlate with Cfs, and not with some other neural state?"

There are two ways of meeting an explanatory request "Why is it the case that *p*?" The first is to produce a *correct answer* to the question—that is, to provide an explanation of why *p* is the case. But it may be that *p* is false and there is no correct answer to "Why *p*?" If someone asks you, "Why did Brutus stab Caesar?" you may be able to come up with a correct answer and meet the explanatory request. In contrast, should a misinformed person ask you, "Why did Brutus poison Caesar?" you cannot provide him with an explanation, for there is none. Rather, you will have to disabuse him of the idea that Brutus poisoned Caesar. The question "Why *p*?" *presupposes* that *p* is the case, and when the presupposition is false, the question has no correct answer, and there is here nothing to explain. Similar remarks apply to "How *p*?" "Where *p*?" and so on. If a child asks you, "How can Santa visit so many millions of homes in one night?" you have to tell her that there is no Santa and he does not visit any homes. You may have to break her heart, but that is the only proper response—that is, if truth matters in this context.

The same goes for the question "Why does pain correlate with brain state Cfs, not with another neural state?" This question presupposes that pain correlates with Cfs, a supposition that is false—if it is indeed the case that pain = Cfs. Pain does not "correlate" with Cfs any more than pain "correlates" with pain. This means that given an identity reduction of mentality—that is, the psychoneural identity theory—it is improper to ask for an explanation of why psychoneural correlations hold. Ned Block and Robert Stalnaker put the point nicely:

> If we believe that heat is correlated with but not identical with molecular kinetic energy, we should regard as legitimate the question of why the correlation exists and what its mechanism is. But once we realize that heat *is* molecular kinetic energy, questions like this can be seen as wrongheaded.[21]

The same goes for pain and neural state Cfs: If we accept "pain = Cfs," the need to explain why pain correlates with this neural state, or why pain occurs iff Cfs occurs, simply vanishes.

What then of the explanatory gap between consciousness and the brain, between pain and Cfs? Again, the identity reductionist should respond: There is no such gap to be closed, and to think that such a gap exists is a mistake—it is the false assumption on which the supposed problem of an explanatory gap

rests. You need two things to have a gap. If identity reduction goes through, that will show that there is no gap, and never was. There is no explanatory gap between pain and Cfs any more than there is one between heat and mean molecular energy, or between water and H_2O.

So the identity reduction of consciousness can handle the explanatory gap, not by closing it but by banishing it out of existence. Clearly, this is a perfectly good way of dealing with the explanatory gap.

One critical question remains, however: Where do we get these psychoneural identities, like "pain = Cfs"? They would be handy things to have: They would let us deal with the explanatory gap problem and help seal the case for physicalism. Their availability, however, has to be shown on independent grounds; it would be question-begging to argue that we earn our entitlement to them simply from their supposed ability to handle the explanatory gap and solve the hard problem. They have this ability only if they are true, and we are entitled to use them to close the explanatory gap only if we have good reason to think they are true. In an earlier chapter (chapter 4), we reviewed three principal arguments for psychoneural identities—the argument from simplicity, the explanatory argument, and the argument from mental causation—and found each seriously wanting. So the main question for the identity solution to the explanatory gap is whether compelling arguments can be produced for psychoneural identities.

Functional Analysis and Reductive Explanation

As we just saw, identity reduction does not give us a reductive explanation of consciousness in terms of neural states; what it does is show that there is no need for such an explanation. Let us now see how a real psychoneural reductive explanation might be formulated. The key to such an explanation is a functional analysis, or functional characterization, of conscious states. Let us suppose that being in pain can be functionally analyzed as follows:

> x is in pain $=_{\text{def}}$ x is in some state P such that P is caused to instantiate by tissue damage, and the instantiation of P causes x to emit aversive behavior.

In a population of interest to us, say, humans, suppose that Cfs is the neural realizer of pain as defined—that is, Cfs is the state that is caused to instantiate by tissue damage and that causes aversive behavior. Given this, we can say

that pain in humans has been *functionally reduced* to Cfs. Thus, functional reduction is another mode of reduction, in addition to identity reduction.

Now consider the following derivation:

(δ) Jones is in Cfs state.
In Jones and organisms like Jones (that is, humans), a Cfs state is caused to instantiate by tissue damage, and it in turn causes winces, groans, and aversive behavior.
To be in pain = $_{\text{def}}$ to be in a state that is caused by tissue damage and that in turn causes aversive behavior.
Therefore, Jones is in pain.

The derivation is valid, and it is plausible to view it as a reductive explanation of why Jones is in pain in terms of her being in a certain neural state. It logically derives a fact about Jones's consciousness from facts about her neural states (the first line), including a neural law (the second line). Note that the third line is a definition, an a priori conceptual truth, not an empirical truth about pain.

Note how (δ) differs from (α) and (β), making use of a psychoneural law and psychoneural identity, respectively. Unlike (α), which appeals to an empirical psychoneural correlation law to make the transition from the neural to the mental, (δ) invokes a definition, in its third line, to make the transition. As may be recalled, it was the use of an unexplained psychoneural correlation law as a premise that doomed (α) as a reductive explanation. The third line of (δ) is not a fact about pains; if it is about anything, it is about the meaning of the word "pain," or the concept of pain. This means that (δ) passes the emergentist's question: "Can we know, or predict, that Jones will be in pain solely on the basis of knowledge of neural-biological facts?" If we have (δ) available, we can say yes; (δ) shows exactly how such knowledge or prediction is possible.

Next, compare (δ) with (β): As we saw, in (β) the conclusion "Jones is in pain" is only a rewrite of "Jones is in neural state Cfs," via the identity "pain = Cfs." This involves no laws, whereas we would expect explanations of events and facts to make use of laws. In contrast, the derivation (δ) makes essential use of a law in its second line; in fact, it is a causal law, and it underwrites the derivational transition from Jones's neural state to her pain. The functional definition of pain is, of course, also crucial, but without the law, the derivation does not go through.

A functional reduction of pain, with Cfs as its realizer, can also generate a reductive explanation of why pain correlates with Cfs. This can be seen in the following derivation:

(ε) *x* is in Cfs state.

In *x* (and systems like *x*), a Cfs state is caused to instantiate by tissue damage, and an instantiation of Cfs state causes *x* to emit aversive behavior.

Being in pain = $_{\text{def}}$ being in a state that is caused to instantiate by tissue damage and that causes aversive behavior.

Therefore, *x* is in pain.

Therefore, if *x* is in Cfs state, *x* is in pain.[22]

This seems like a perfectly good explanation of why Cfs correlates with pain, in *x* and systems like *x*.[23]

But all this is contingent on the availability of functional definitions of mental states, especially states of phenomenal consciousness, or qualia. At this point the situation with the present approach is similar to the situation facing psychoneural identity reduction. Both can handle the explanatory gap problem, each in its own way. Identity reduction promises to make the explanatory gap disappear. Functional reduction meets the explanatory demand head-on, providing reductive explanations of psychoneural correlations. But just as identity reduction must come up with psychoneural identities to turn its promise into reality, functional reduction will remain an empty schema if consciousness properties, such as pain and visual experience of green, turn out to resist functional analysis. Before we turn to these questions in the final section, let us pause to consider consciousness in the context of neuroscience.

Consciousness and Brain Science

There are three connected but separable questions about brain science and consciousness. They are:

1. Can consciousness be given a scientific account—presumably in neural-behavioral-computational terms? That is, is consciousness a proper and appropriate *explanandum* in science?
2. Can, or does, consciousness have a theoretical/causal/explanatory role in neural-behavioral science? That is, can consciousness play a role as an *explanans*, something that has explanatory power and efficacy, in neural-behavioral science?
3. Can consciousness be studied scientifically? Can it be investigated by the current methodologies in neural-behavioral science?

We have in effect answered the first question about the possibility of a scientific account of conscious states. In the preceding section, we saw that if a conscious state can be given a functional analysis, in terms of physical stimulus input and behavioral output, it is in principle possible to provide a reductive neural explanation for the occurrences of the conscious state (assuming that the relevant neural realizers have been identified). We saw in concrete terms how pain occurrences in a person could be derived from neural facts about the person, with the aid of a functional characterization of pain.

This means that the answer to the first question is yes—*if* conscious states are functionally analyzable and *if* we have been able to identify their neural realizers (in populations of interest). These are two big "ifs." The second "if" concerns progress of research in cognitive neuroscience, and it assumes that the first "if" has been satisfied. If a conscious state is not functionally analyzable, it isn't something that can have realizers, neural or otherwise. So the first "if" is the philosophically important "if."

Most philosophers will agree that a standard sort of functional definition of pain, like the one adverted to in the preceding section, in terms of stimulus input and behavioral output, does not work. It does not do justice to the qualitative aspect of pain; pain as a quale cannot be functionalized. Philosophers who now advocate a functional approach to qualia, however, take a different tack, and that is qualia representationalism (see chapter 9). On this view, qualia are essentially representational, as are all other conscious states, and representation is fundamentally a functional concept. Let us see how this is supposed to work for pain. An experience of pain is a representational state, a state with a representational object or content. It is important to keep in mind that pain, or being painful, is not a property, or quality, of the experience; rather, pain is what the experience represents—it is the object, or content, the pain experience represents.

So, then, what do pain experiences represent? According to Christopher Hill, such experiences represent bodily disturbances, such as a burned finger, a scraped knee, or a broken arm.[24] Since pains are the objects represented by pain experiences, it follows that pains *are* bodily disturbances. When we are aware of a pain, we are almost always aware of its location—in the elbow, in the thumb, and so on—and we are made aware that there is something wrong, at any rate not quite right, going on at that location. In other words, our pain experiences seem directed at bodily disturbances, either at the body's periphery or deeper inside. This means that bodily disturbances are the intentional objects, or contents, of our pain experiences, and there is ample reason to think that it is the biological function of pain experiences to represent bodily distur-

bances.[25] This allows the possibility that there is a pain experience but no actual bodily disturbance represented by it, like in cases of "phantom limb pain"; it is simply a case of misrepresentation (see chapter 8). In any case, all there is to pain experiences, on this account, is for the experiences to represent bodily disturbances, and that exhausts the real nature, or essence, of these experiences as pain.

If such an account is found to work, that would open the possibility of a broadly functional account of pain, or pain experience. (The reader is encouraged to work out, in schematic form, how such an account may be formulated.) If a similar representational account of consciousness works in general, the essential nature of conscious states will be fully captured by their representational properties. There is a general agreement, or presumption, that representation is more physicalistically tractable than phenomenal consciousness, or qualia, as usually conceived. However, as has been noted in other contexts, unless you are irrevocably and antecedently committed to physicalism, the plausibility of consciousness representationalism should be assessed on its own merits. This comment applies to Hill's bodily disturbance theory of pain. One might wonder, for example, whether this view of pain does full justice to the experiential aspect of pain. On the representational account, when pain occurs, is there anyone who is actually hurting? Why should my representing a bodily disturbance, say, a cut in my finger, be experienced as painful? The reader may also ponder the "cross-wired brain" case (discussed in chapter 6), in which the "pain box" and the "itch box" exchange their input and output channels.

As for our second question, concerning the possible theoretical-explanatory role consciousness might have in the behavioral and brain sciences, we can approach it by considering the causal status of consciousness. First, though, we should remember that we are not concerned here with access consciousness; for the concept of access consciousness is fully functional, characterizable by its role in the cognitive economy of a cognitive-psychological subject, and hence does not present us with any additional philosophical issues. So qualia, or phenomenal consciousness, are our present focus. Our question, then, is this: Can qualia play an explanatory role in the brain and behavioral sciences? The companion causal question is: Do qualia have causal powers to affect behavioral or neural events?

Given the causal closure principle for the physical domain, the answer to this question has to be: Highly unlikely—unless qualia are reductively identified with neural states. What we saw in our discussion of mental causation (chapter 7) was that if any object or event were to exercise its causal powers in the physical domain, it must be part of that domain, or be reducible to it; the

physical domain does not tolerate causal influences from outside. But beyond such overarching doctrines as the principle of physical causal closure, there are more intuitive considerations for thinking that qualitative conscious states are epiphenomenal—or, at least, that this is the way many of us regard it.

Suppose a brain researcher comes across a neural event for which she is unable to identify a physical-biological cause—for which she has difficulty providing a physical-biological explanation. What is the chance that she will decide to explore the possibility that some purely psychic event, a nonphysical conscious event, might be the cause of this unexplained neural event? Would a working neuroscience researcher ever look to the realm of nonphysical consciousness for causes of neural-biological-behavioral events? We may be pretty sure that there is no such chance. In the first place, how would she show that a specific conscious event is its cause? Surely the conscious event has a neural substrate, or correlate. Wouldn't this neural substrate always be a better candidate as the sought-after cause? She may not know what the neutral substrate is; she doesn't have a biological-physical description of it. But she is entitled, it seems, to believe that such a neural substrate must exist, and her research time will be better spent on focusing on identifying it in physical-biological terms and studying its properties.

It seems, then, that in practice scientists working in brain science are guided by something like the following principle:

> *The causal-explanatory closure of the neural-physical domain*. If a neural-biological event has a cause—or causal explanation—it has a neural-physical cause and a neural-physical explanation.

That is to say, the neural-physical domain is causally and explanatorily self-sufficient. And this naturally leads to the following principle:

> *Methodological qualia epiphenomenalism*. Qualia, unless they are reducible to physical-neural properties, are to be treated as epiphenomenal. They should not be invoked as causes of, or in causal explanations of, neural-physical-behavioral events.

What is being suggested, then, is that in practice qualitative consciousness is treated by scientists as epiphenomenal with respect to the physical domain. And this gains some empirical support from the fact that much of our cognitive processing goes on at the subpersonal, or unconscious, level, and that qualia are scarcely mentioned in serious theories of cognitive science. This is

not to say that such concepts as awareness, attention, and signal detection play no role in cognitive science; these are concepts of access consciousness, not phenomenal consciousness. So our answer to the second question is no: Phenomenal consciousness, being epiphenomenal, has no role to play as a theoretical-explanatory concept in neuroscience.

We now turn to our third question: Can phenomenal consciousness be investigated scientifically? Can there be a scientific theory of qualia? Our discussion of the second question argues powerfully for a negative answer: Qualia are not proper objects of scientific study. Here is why. Qualia are epiphenomenal; they cause no effects in the physical domain. If so, how can they even be observed? How can their presence be known to the investigator? There can be no instrument to detect their occurrence and identify them. Qualia cannot register on any measuring instruments because they have no power to affect physical objects or processes. No one thinks the brain scientist can "directly" observe a subject's conscious state, phenomenal or nonphenomenal; direct observation of a conscious state requires experiencing it, and the scientific observer of course is not experiencing the subject's conscious states. The brain scientist relies heavily on the subject's verbal reports in attributing conscious states to her. However, if qualia are epiphenomenal, they have no role in causing verbal reports, which involve physical processes, like vibrations of the vocal cords. If so, how could verbal reports be *evidence* for the presence or absence of qualia? The practice of using verbal reports to determine what conscious states are occurring presupposes that conscious states are causes of the reports.

What about the pulsating yellow and orange images we see on an *f*MRI monitor? Don't they show that the subject is seeing green, feeling agitated, recalling a traumatic event from the past, and so on? Aren't these images caused by visual qualia, qualia associated with emotions, and the like? Well, no. They are caused by physical processes, in fact, patterns of blood flow. Qualia cannot cause these physical processes; they are epiphenomenal. You might retort: As you said, we believe that a pervasive system of mind-brain correlations exist, and that each conscious state, of any kind, has a neural correlate. Doesn't this mean that we can ascertain the occurrences of conscious states by noting the occurrences of neural states?

Again, we have a predictable reply: If qualia are truly epiphenomenal, a serious question arises as to how we could establish these qualia-brain correlations to begin with. We usually establish a correlation, "M occurs iff N occurs," by observing the co-occurrences of M and N. But if either M or N is epiphenomenal, its occurrence cannot be observed, and hence we are not in a position to confirm, or disconfirm, the posited correlation between M and N.

Brain scientists do not rely solely on the subjects' verbal reports. They heed the subjects' behaviors to supplement and corroborate verbal reports. But this practice, too, presupposes that conscious states are the causes of behaviors. Again, if qualia are epiphenomena, they can play no role in behavior production, and behaviors cannot be evidence for the presence or absence of qualia.

One last question: If phenomenal consciousness is indeed an epiphenomenon, with no causal efficacy at all, how could it have evolved? Mustn't every evolved feature of organisms have causal efficacy? The standard answer to this question is no—the evolved feature may simply be an "unintended" collateral feature of another feature that was selected by natural selection—a so-called spandrel effect.[26] Polar bears have thick and heavy fur. The heaviness of the fur has evolved, but that was only a side effect of their thickness, which is the feature selected for.[27] It is brain functions that have causal effects on adaptive behaviors of organisms, and it is the neural mechanisms that perform these functions that are selected by natural selection. Phenomenal consciousness can be regarded as an unintended side effect of the evolution of the neural systems in higher organisms. Another point can be added: If qualia have adaptive values, it isn't their intrinsic qualities, or qualia as intrinsic qualities, that are valuable; it is qualia differences and similarities, namely, sensory discriminations based on qualia differences, that have behavioral consequences (think about traffic lights) and confer adaptive advantages on organisms. Qualia as intrinsic qualities remain epiphenomenal.

Scientific research on consciousness has been flourishing during the past several decades, and we can only expect it to continue to thrive, bringing to us new insights into how our minds work.[28] Doesn't this mean that methodological epiphenomenalism is not constraining the research practices of brain scientists after all, and that practicing brain scientists do believe in the causal efficacy of consciousness? Perhaps so, but, again, keep in mind that the epiphenomenalism we have been discussing concerns only phenomenal consciousness, the felt and experienced qualities of conscious states, and not states of consciousness that fall under access consciousness. Whatever it is that consciousness researchers are investigating and theorizing about, it can't be phenomenal consciousness. Or so goes the epiphenomenalist story.[29]

In this section, we have put the consequences of qualia epiphenomenalism in stark and uncompromising terms, in part as a challenge to the reader to ponder and reflect on them. These issues are important not only to the brain and behavioral scientists but to all of us on a personal level. How could they not be if, as Wilfrid Sellars put it, qualia are what make life worth living? Or if,

as Ivan Pavlov said, at the end of the day, our "psychic life" is the only thing that matters to us?

What Mary, the Super Vision Scientist, Didn't Know

In a paper published in 1982, Frank Jackson presented a thought experiment featuring Mary, a talented vision scientist, on the basis of which he formulated a much-debated antiphysicalist argument. To see how the argument runs, we can do no better than quote a paragraph from this paper:

> Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room via a black and white television monitor. She specializes in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like "red," "blue," and so on. She discovers, for example, just which wavelength combinations from the sky stimulate the retina, and exactly how this produces *via* the central nervous system the contraction of the vocal chords [*sic*] and expulsion of air from the lungs that results in the uttering of the sentence "The sky is blue." . . .
>
> What will happen when Mary is released from her black and white room or is given a color television monitor? Will she *learn* anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had *all* the physical information. *Ergo* there is more to have than that, and Physicalism is false.[30]

What is "physical" information? For the purposes of his argument, Jackson takes information from the physical, chemical, and biological sciences to be physical information. And his formulation of physicalism is this:

> *Physicalism*. All (correct) information is physical information.

With this in hand, we can set out Jackson's antiphysicalist argument like this:

1. Before her release from the black-and-white room, Mary had all the physical information about human vision.

2. When she first gazes at a ripe tomato after her release, she gains new information—she learns something new about human vision.
3. Hence, the new information she gains is not physical information.
4. Hence, there is information other than physical information, and physicalism is false.

This is the celebrated "knowledge argument," and we can see why it is so called. First of all, Jackson's formulation of physicalism is epistemic: It concerns what types of information, or knowledge, there are about the world. He is not explicit about what he means by information, but it seems that we can safely use "information" and "knowledge" interchangeably in this context. For contrast, we can state a metaphysical thesis of physicalism, something that arguably is more central to philosophical concerns:

Metaphysical Physicalism. All facts are physical facts.

And the crucial premise of the argument is that on her release Mary gains new information, that is, new knowledge that is nonphysical.

There are two main questions about this argument. First, since the argument obviously is valid, are the premises of the argument true? Second, if the argument is correct, does it show anything about metaphysical physicalism? Most critics of the argument have focused on the second premise, the claim that Mary gains new knowledge when she first looks at a ripe tomato. According to an influential reply, what Mary gains is not propositional knowledge, or knowledge of a fact, but a set of abilities—abilities to recognize red and other colors, to imagine colors, to make color similarity and difference judgments by looking, and the like. Perhaps she gains a new "recognitional concept," a disposition or ability to recognize and classify objects, events, and phenomena through perceptual discrimination.[31] Knowing what it's like to see red is a case of knowing-how, not knowing-that. This is called the Ability Hypothesis.[32]

It may be conceded that on her release from her monochromatic environment, Mary certainly gains these abilities, but that does not preclude her also gaining propositional knowledge, knowledge of new facts about how things look to her and other people. Is there any reason to think that this is not the case? If she acquires new propositional knowledge, there must be a proposition that she comes to know. When Mary looks at a red tomato for the first time, what exactly is the proposition that she comes to know? She says to herself, "Isn't that interesting! So this is what red looks like." But can we put her new knowledge in propositional form, that is, in a declarative sentence? It

seems that the best we can do is something like "Red looks like *this*," where "this" is used as a demonstrative pointing to the red tomato. It's hard to see how we could avoid using demonstratives in formulating a proposition representing her new knowledge.

Is this a problem? A critic may charge that this shows something odd and unusual about the supposed propositional knowledge Mary gains; if it is indeed a piece of propositional knowledge, that is, knowledge of a fact, shouldn't it be possible to express its content without using a demonstrative, which could be understood only in relation to the person using it and the particular context in which it is used? Any objective information must be expressible in a sentence free of demonstratives ("this," "that," etc.) and indexicals ("I," "here," "now," etc.).

It seems, though, that friends of the knowledge argument have a ready reply: On the contrary, the fact that demonstratives must be used only goes to show that the knowledge gained is not physical knowledge. Physical knowledge is objective in the sense that it is neutral with respect to "points of view"—that is, the perspective of an observer or experiencer. In contrast, experience is always experienced from a single point of view, namely the experiencer's, and it is no surprise that Mary's new knowledge must be expressed as "Red looks like *this*"—"like *this*" to Mary. (Recall Nagel and his bats.) Thus, the essentiality of demonstratives in expressing contents of Mary's knowledge is a reflection of the subjective character of her knowledge—the fact that her knowledge is not physical. This only provides more support for the knowledge argument.[33]

Some may protest that Jackson's Mary is not a real possibility—that during her confinement, she could have dreamed in color, that she might have accidentally rubbed her eyes in such way as to cause color experiences, that she could have directly stimulated her visual cortex to experience color, and so on. These are all possible, but the reply misses the point. Mary is a thought experiment, and Jackson is free to set it up any way that suits him. All he needs to suppose is that none of these possible situations occurred, and that Mary in fact had no color experience before her release. If you say she did, that would be changing the example—and the subject.

It seems clear that when Mary exits from her confinement into the outside world, an important cognitive change occurs to her. That much we must all accept. The only issue is how this cognitive change is best understood. Those who are antecedently committed to physicalism would try to describe it in a way consistent with physicalism; the Ability Hypothesis is one such attempt. Whether this and other physicalist responses are at all persuasive is a question that is still open.[34]

Let us turn to our second question—what the knowledge argument might show about metaphysical physicalism, the thesis that all facts are physical facts. It is pretty obvious what tack the physicalist would take on this question. She would argue that Mary's new knowledge is not of a new nonphysical fact, but of an old physical fact in a new guise. Consider an analogy: Ancient Greeks knew that water extinguished fire but did not know that H_2O extinguished fire. They had no knowledge whose content is expressed by using the concept H_2O. When we learned that water = H_2O, we also learned that H_2O extinguished fire. But this is not knowledge of a new fact; the fact that H_2O extinguishes fire is the same fact as the fact that water extinguishes fire. An old fact that we had known for centuries was given a new description; we may speak of new knowledge if we like, but what is important is to see that no new fact came to be known. Similarly, in the case of Mary, the fact that tomatoes look a certain way to her is just the fact that the surface reflectance property of tomatoes is such and such. Now, on account of her direct visual experience, she can express this fact in a new way: "Fancy that! Tomatoes look like this!" It would follow, then, that even if the knowledge argument succeeds as a refutation of Jackson's epistemic physicalism, it may have no adverse effect on metaphysical physicalism.

How plausible is this brief in behalf of metaphysical physicalism? In pondering this question, the reader should heed its similarity with psychoneural identity theory, which, too, involves the claim that the fact stated by "I am in pain" is identical with the fact stated by "My C-fibers are stimulated," and that here is one fact under two descriptions, not two distinct facts. Thus, arguments pro and con in regard to psychoneural identity theory may well be relevant here as well.

The Limits of Physicalism

In an earlier section, we saw how two approaches to consciousness, psychoneural identity reduction and functional reduction, could deal with the explanatory gap and the hard problem of consciousness. It is also easily seen how they could deal with the problem of mental causation and fight off the threat of epiphenomenalism. If psychoneural identity theory can be upheld and we are in a position to affirm "pain = C-fiber stimulation," "consciousness = pyramidal cell activity," and the like, there will be no special problem about how pain, or consciousness, can cause, and be caused by, other events. For pain will have exactly the causal properties of C-fiber stimulation, and similarly for consciousness and pyramidal cell activity. Under the identity ap-

proach, all causal actions take place in the physical domain and the mental is part of that domain.

Let us see how mental causation works out under functional analysis. Suppose pain could be given a functional analysis in terms of its causal role—as the causal intermediary between pain input (tissue damage) and pain output (aversive behavior). That is, to be in pain is to be in some state that plays this causal role. If so, when you are in pain, you must be in some state that "realizes" pain—that is, in some state that plays the causal role distinctive of pain. In this particular instance, let us suppose, you are in pain in virtue of being in the state of C-fiber stimulation (Cfs), which realizes pain in you and other humans. So here is an instance of pain and an instance of Cfs. How are these two instances, or token events, related to each other? The answer that they are one and the same event is all but compelling. To be in pain *is* to be in a state that plays pain's causal role. Hence, for you to be in pain on this occasion is for you to be in a state that plays pain's causal role, and Cfs is the state you are in that plays pain's causal role. It evidently follows that for you to be in pain on this occasion *is* for you to be in Cfs state on this occasion. That is, your pain instance is identical with your Cfs instance. It further follows that your pain instance and Cfs instance have the same causal powers, since they are one and the same. This solves the problem of mental causation for your pain instances. This idea generalizes to all other cases of mental events and states: Each instance of a mental property has the causal powers of the instance of its physical realizer.[35] The threat of epiphenomenalism has been vanquished.

To take stock: Either psychoneural identities or functional analyses of mental states, each in its own way, can deal with mental causation and the explanatory gap. These two approaches can be considered two ways in which mentality can be physically reduced: The first does so by identifying mental states with neural-physical states; the second accomplishes reduction through functional analysis of mental states. If we are to avoid epiphenomenalism, physical reductionism is the only alternative; in one way or another, we must bring the mental within the physical fold. That much is a direct consequence of the principle of physical causal closure (see chapter 7). However, our willingness to countenance reductionism does not in itself show that reductionism is true—that is, that either kind of reduction is really possible for the mental. The reducibility of the mental has to be shown on independent grounds. If, despite our willingness to cast our lot with reductionism, the mental turns out to be physically irreducible, epiphenomenalism cannot be avoided. So is the mental reducible, and if so, how?

In considering these questions, one pitfall to avoid is the tendency to think that the mental in its entirety must be either reducible or irreducible. It may well be that some mental properties are reducible while others are not. For the would-be reductionist, it would be better if the former were to outnumber the latter. The main point to remember is that the reductionist project need not be a total success or a total failure. The more it succeeds, the more we succeed in saving mentality from epiphenomenalism; the less successful it is, the fewer the mental properties we can save from causal impotence.

You may recall a broad classification of mental phenomena into two kinds (see chapter 1): phenomenal mental events, or experiences, with sensory and qualitative characters, like bodily sensations, seeing yellow, smelling ammonia, and the rest, on the one hand; and intentional-cognitive states (or propositional attitudes), like belief, desire, intention, thought, and the like. The former are states with "qualia"; there is a "what it is like" quality to having them or being in them. The latter have propositional contents expressed by subordinate that-clauses ("Bill believes that there are lions in Africa," "Ann hopes winter will be mild this year"). You may recall the question of what events and states in these two classes have in common that makes all of them mental. It may well be that physical reducibility is not among the shared properties of these two mental categories.

There is a view on the current scene according to which the former, states with qualia, are irreducible, whereas the latter, intentional-cognitive states, are so reducible.[36] Let us take up the second class of mental events first. Why should we think beliefs, desires, and such are reducible, and if so, according to which model of reduction? It would seem that these states cannot be reduced by identity reduction—that is, it is not possible to identify them with neural-physical states. The reason for this is the old and familiar nemesis of reductionism, the multiple realizability of these states (see chapters 4 and 5). However, functional reduction, or reduction via functional analysis, can accommodate multiple realization, because functionally defined states or properties can have multiple realizers. But can these states then be functionally analyzable or definable?

To reduce a property functionally, the property must first be functionally analyzed or defined. This is the required conceptual preliminary. After a property has been functionalized, it is the job of science to discover its realizers (in populations of interest). So the question for us is this: Can intentional-cognitive states, like beliefs and desires, be given functional characterization? Can we analyze belief as an internal state defined in terms of its serving as a causal intermediary between inputs and outputs?

It has to be admitted that, as critics of functional reduction would argue, no one has yet produced a complete functional definition or analysis of belief and that none is in sight. However, there are reasons for thinking that belief and other intentional-cognitive states are functionally conceived states—that is, states understood in terms of their "job descriptions." We consider here two such reasons. First, there seems ample ground for believing that intentional-cognitive states are supervenient on the physical-behavioral properties of creatures. Consider the "zombies"—the supposedly conceivable creatures who are just like us both in internal compositional-structural detail and in the functional organization of sensory input and behavioral output but who lack experiences—that is, they have no phenomenal consciousness; there is nothing it is like to be a zombie. Whether such creatures could exist is a question we need not address here. Our immediate question is whether zombies have intentional-cognitive states, and a strong case can be made for saying that they must have such states. To begin, the zombies are indistinguishable from us, and our fellow humans, behaviorally and physically. If that is the case, we must attribute to them a capacity for speech. They emit noises that sound exactly like English sentences (actually, quite a few zombies will speak Chinese!), and they apparently communicate among themselves by exchanging these noises and are able to coordinate their activities just as we do. Moreover, if zombies are among us, they will talk to us and we can understand what they are saying. And apparently they understand us when we talk to them. They read newspapers, surf the Internet, and watch television. Remember: These zombies are behaviorally indistinguishable from humans. Given all this, it would be incoherent to deny that they are language users. Zombies have speech, just like us.

A language user, by definition, is capable of performing speech acts. Making assertions is a fundamental speech act, and any creature with speech must be able to make utterances and thereby make assertions. Further, to utter "Snow is white" to make an assertion is to express the *belief* that snow is white. Consider other speech acts, like asking questions and issuing commands. To ask "Is snow white?" is to express a *desire* to be told whether snow is white. To command "Please shut the window" is to express the *desire* that the window be shut and the *belief* that the window is not now shut. It is not conceptually possible to concede that zombies are language users and then refuse to attribute beliefs and desires to them. Once these states are attributed to the zombies and given the assumption that they are behaviorally indistinguishable from us, we must also recognize them as full-fledged agents. Thus, we seem to be driven to the conclusion that belief, desire, intention, agency, and the rest are

supervenient on the physical-behavioral aspects of creatures and that these states cannot go beyond what can be captured in physical-behavioral terms. This contrasts with the case of qualia supervenience; as we saw earlier in this chapter, there are reasons to be skeptical about the supervenience of qualia on physical-behavioral properties.

Second, suppose we are asked to design and build a device that detects shapes and colors of objects around it (perception), processes and stores the information it has gained (information processing, memory, knowledge), and uses it to guide its behavior (action). If that is our assignment, we would know how to go about executing it; there probably already are robots with such capabilities in limited form. We know how to proceed with the design of such a machine because processes and states like perception, memory, information processing, and using information to guide behavior and action are defined by job descriptions. That is, these concepts are functional concepts. A device, or creature, that has the capacity to do certain specified work under specified conditions is ipso facto a system that perceives, processes, and stores information, makes inferences, and so on. The main point to note here is that these intentional states and processes are tied to having capacities of certain kinds—capacities to interact and cope with the environment. The only difference between such states of our simple machine and real-life intentional-cognitive states is that the causal tasks associated with the former are exactly specified and limited in scope, whereas those associated with the latter are less precisely defined and, more important, open-ended.

It may be true, as critics of functional reduction of intentional-cognitive states have argued, that we will never have complete functional specifications of intentional states like belief, desire, and intention. But that is only because, as just noted, the causal tasks involved with belief are open-ended and perhaps essentially so. It does not show that these are not functional, task-oriented states; as far as supervenience holds, there cannot be extra factors beyond functional-behavioral facts that define or constitute them. More important, it is not necessary to have full and complete characterizations of these states before scientific research can begin to identify their physical realizers—the neural mechanisms that do the causal work so far specified. The fact is that these intentional states are multitask states; their cores may be fairly easily identifiable, but they may have no clear definitional boundaries. Obviously, belief must be closely connected to a system's speech center and feed into its inference- and decision-making modules. What else do beliefs do? That can remain an open question. Moreover, there probably is no clear line between those functions of beliefs that ought to be part of the concept of belief and those that

are contingently, though nomologically, connected to beliefs. As scientific research makes progress, we will probably keep adding to and subtracting from the initial job descriptions of these states; that is one way in which our concepts change and evolve.

Compare this with the situation involving qualia—or the phenomenal, qualitative states of consciousness. Suppose that you are now given an assignment to design a "pain box," a device that can be implanted in your robot that not only will detect damage to its body and trigger appropriate avoidance behavior but also will enable the robot to experience the sensation of pain when it is activated. Building a damage detector is an engineering problem, and our engineers, we may presume, know how to go about designing such a device. But what about designing a robot that can experience pain? It seems clear that even the best and brightest engineers would not know where to begin. What would you need to do to make it a pain box rather than an itch box, and how would you know you have succeeded? The functional aspect of pain can be designed and engineered into a system. But the qualitative aspect of pain, or pain as a quale, seems like a wholly different game. The only way we know how to build a pain-experiencing system may well be to make an exact replica of a system known to have the capacity to experience pain—that is, to make a replica of a human or animal brain. Even if we could make replicas of a brain, that still would not give us an understanding of how experiences of pain and other qualia arise from the electrochemical processes of the brain.

Some philosophers have argued that zombies (without inner experiences) are metaphysically possible and therefore that the qualitative states of consciousness are beyond the reach of physicalism. The zombie hypothesis has been controversial, and we do not need the zombies to see that qualia are not functionally definable. All we need is the possibility of qualia inversion—for example, visual spectrum inversion. All we need is the logical possibility of a world that is physically indiscernible from this world but in which people's color spectrum is inverted in relation to ours. The conceivability and possibility of such a world should be less controversial than that of a zombie world.[37] People in the color-inverted world behave exactly as we do; their functional-behavioral properties are exactly the same as ours, and yet their color experiences are different.[38] If such a world is possible, it follows that color qualia are not functionally definable and hence not functionally reducible. Pain certainly has a function, the important biological function of separating us from sources of harm, teaching us to avoid potentially noxious stimuli, and so forth. However, its function is not, it may be argued, what makes pain pain, or what constitutes pain; rather, it is the way it feels—nothing can be a pain unless it *hurts*.

If qualia resist functionalization, as they seem to, they cannot be reduced by functional analysis—that is, functional reduction doesn't work for them. What about the prospect of an identity reduction of qualia? Earlier (chapter 4), we discussed several positive arguments for psychoneural identities and found all of them seriously defective or incomplete. The argument from simplicity, of the sort J. C. C. Smart originally appealed to, does not have enough weight behind it to be convincing; simplicity-based arguments seldom do when what is at issue is the *truth* of the doctrine being defended. No one has shown why the supposedly simplest hypothesis must be the true one. We saw how the two explanatory arguments on the scene fall short of the mark; these arguments invoke something like the rule of inference to the best explanation, but we saw how these arguments misapply this rule (and the rule itself is not uncontroversial). The causal argument seems to work better, but it does not go the full distance; in effect, it only shows the conditional proposition that if mental causation is to be saved, the mental must be brought into the physical domain—that is, physically reduced. And this exactly is the issue we are now grappling with.[39] We have to conclude that an identity reduction of qualia is no more promising than their functional reduction and that qualia epiphenomenalism looms as a real threat.[40]

We should note, however, that saving intentional-cognitive states from epiphenomenalism is not a small accomplishment. In saving them, we save ourselves as agents and cognizers, for cognition and agency are located in the realm of the intentional-cognitive, not the phenomenal. Recall Fodor's lament about the possible loss of mental causation:

> If it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying . . . , if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world.[41]

Three items are on Fodor's wish list: wanting, itching, and believing. We can reassure Fodor that his world is not coming to an end, at least not completely. We can save wanting and believing; that is, we can save agency and cognition. Two out of three isn't bad!

But what about itching? There are reductive approaches to consciousness that attempt to reduce it to intentional-representational states. Two such approaches, the higher-order perception/thought theory and qualia representationalism, were reviewed earlier (see chapter 9). If these theories work and we

can reduce qualia to intentional-representational states, these latter states could in turn be functionalized, and that would yield a solution to both the mental causation problem and the explanatory gap problem. Just because these approaches to qualia would do something nice for us, perhaps something very important, that is not a reason to think that they must work. They must first be shown to be correct approaches, and we have seen some serious difficulties for both, though the representational approaches are very much alive. It is fair to say that qualia representationalism is currently the leading physicalist approach to phenomenal consciousness.

Returning to the model of functional reduction, we can go a little more distance toward saving qualia. Begin with an analogy: traffic lights. Everywhere in the world, red means "stop," green means "go," and yellow means "slow down." But this is merely a conventional arrangement; as far as traffic management goes, we could do just as well with a system whereby red means "slow down," green means "stop," and yellow means "go"—or any permutation thereof. What is important is our ability to *discriminate* among these colors; the colors themselves do not matter. The same holds for qualia. You and your spectrum-inverted friend will do equally well in coping with traffic lights, in picking tomatoes out of mounds of lettuce, in using color words to report visual experiences, and in learning about what is out there in your respective surroundings. It is qualia differences and similarities, not qualia as intrinsic qualities, that matter to our perception and cognition. That roses look *this way* and irises look *that way* cannot be cognitively relevant as long as roses and irises look relevantly different. Qualia differences and similarities are behaviorally manifest, as we just saw, and this opens the door to their potential functionalization and reduction.

We can conclude, therefore, that qualia are not entirely lost to epiphenomenalism; we can save qualia differences and similarities, if not qualia as intrinsic qualities. So what we may lose to epiphenomenalism, and something for which we cannot solve the explanatory gap problem, is this small mental residue, qualia as intrinsic qualities, untouched and untouchable by physicalism. And that represents the limits of physicalism.

For Further Reading

On the explanatory gap, see Joseph Levine, "Materialism and Qualia: The Explanatory Gap" and "On Leaving Out What It's Like." Levine's *Purple Haze* is his most recent and developed statement on the issues of phenomenal consciousness. See also David J. Chalmers, *The Conscious Mind*. For analysis and

critique, see Ned Block and Robert Stalnaker, "Conceptual Analysis, Dualism, and the Explanatory Gap," and Block, "The Harder Problem of Consciousness." Other discussions include David Papineau, *Thinking About Consciousness*, and John Perry, *Knowledge, Possibility, and Consciousness*. Both Papineau and Perry defend physicalism against well-known objections, like the zombie argument and the knowledge argument. Daniel Stoljar's *Physicalism* is a readable up-to-date survey, analysis, and discussion.

The Waning of Materialism, edited by Robert C. Koons and George Bealer, is a recent anthology of new essays critical of the materialist-physicalist paradigm.

There is a large literature on the knowledge argument. Two collections of essays are worth examining: *There's Something About Mary*, edited by Peter Ludlow et al., and *Phenomenal Concepts and Phenomenal Knowledge*, edited by Torin Alter and Sven Walter.

The Case for Qualia, edited by Edmond Wright, collects recent essays defending qualia against the deflationist-eliminativist stance taken by many contemporary philosophers.

On qualia epiphenomenalism, see Frank Jackson, "Epiphenomenal Qualia," and Jaegwon Kim, *Physicalism, or Something Near Enough*, chapter 6. The latter presents in greater detail the overall picture described in the last section of this chapter. *The Conscious Mind* by David Chalmers presents a similar picture.

Notes

1. This term is due to W. V. Quine.

2. Note that there can be multiple supervenience bases for a mental state. N may be the supervenience of pain for you, but as we have seen with the multiple realizability of mental states (chapter 5), a different neural state may be pain's supervenience base for octopuses, still another for reptiles, and so on.

3. The term "explanatory gap" was introduced by Joseph Levine in his "Materialism and Qualia: The Explanatory Gap." The issue of explaining mind-body supervenience relations is highlighted in Terence Horgan, "From Supervenience to Superdupervenience."

4. This formulation of the question is Ned Block's.

5. William James, *The Principles of Psychology*, p. 647 in the 1981 reprint edition.

6. T. H. Huxley, *Lessons in Elementary Physiology*, p. 202.

7. David Chalmers, *The Conscious Mind*, p. 24.

8. Such scanning devices must ultimately be neural organs. If so, it is at least conceivable that your scanning system gets hooked up with my brain so

that it monitors my first-order mental states, and conversely that my internal scanner is wired to your brain to monitor your first-order states. In this situation, would you be conscious of my mental states, and I of yours? Does this even make sense? If the internal monitoring account of consciousness implies this to be a possible situation, that might be a sign that there is something deeply wrong with the account.

9. Saul Kripke, *Naming and Necessity,* pp. 153–154. The target of Kripke's argument is the identification of pain with C-fiber stimulation; however, his argument applies with equal force against the supervenience of pain on C-fiber stimulation.

10. This is based on Ned Block's "Inverted Earth."

11. Arthur Rimbaud, "Voyelles." The phenomenon of synesthesia, in which a person, for example, hears sounds when she sees motion, makes it easier to imagine inverted sense modalities.

12. For complexities and complications in the supposition of inverted spectra, see C. L. Hardin, *Color for Philosophers*. See also Sydney Shoemaker, "Absent Qualia Are Impossible: A Reply to Block" and "The Inverted Spectrum"; and Michael Tye, "Qualia, Content, and the Inverted Spectrum."

13. This point is discussed in connection with functionalism; see chapter 5.

14. It is consistent to hold the supervenience of qualia on physical properties but deny their supervenience on functional properties. We might, for example, hold that qualia arise out of biological processes and that there is no reason to think that qualia are experienced by an electromechanical system (say, a robot) that is functionally indistinguishable from us.

15. There has been an active and wide-ranging debate over the relationship between conceivability and real possibility. The collection *Conceivability and Possibility*, ed. Tamar Szabo Gendler and John Hawthorne, includes a number of interesting papers on the topic (including a comprehensive introduction).

16. We saw two advocates of this option in the preceding chapter, Daniel Dennett and Georges Rey.

17. Jerry Fodor writes, "If mind/body supervenience goes, the intelligibility of mental causation goes with it," *Psychosemantics*, p. 42. See Terence Horgan, "Supervenient Qualia," for a causal argument for qualia supervenience.

18. David Chalmers, *The Conscious Mind*, p. 43. Emphasis in original.

19. Jerry A. Fodor, "Special Sciences, or the Disunity of Science as a Working Hypothesis," in *Philosophy in Mind: Classical and Contemporary Readings,* ed. David J. Chalmers, p. 131.

20. For more details on scientific explanation, see Carl G. Hempel, *Philosophy of Natural Science*.

21. Ned Block and Robert Stalnaker, "Conceptual Analysis, Dualism, and the Explanatory Gap," p. 24.

22. The derivation of this line is by a logical rule called "conditionalization," whereby a premise is "discharged" by making it the antecedent of an "if . . . then" statement with the last proved conclusion as the consequent.

23. To derive a full "iff" correlation, we also need to derive "*x* is in Cfs state" from "*x* is in pain." The reader might want to try such a derivation.

24. Christopher Hill, *Consciousness*, chapter 6. Hill also offers another physical theory of pain, the somatosensory theory, according to which pains are somatosensory representations of bodily disturbances, though the bodily disturbance theory remains his preferred option. For details and defense of the bodily disturbance account, the reader should turn to Hill's presentation and discussion in his book.

25. There are people who are congenitally incapable of experiencing pain. They have great difficulty coping with their surroundings without injuring themselves, and most of them do not live to adulthood.

26. The term "spandrel effect" was introduced by the evolutionary biologists Stephen Jay Gould and Richard Lewontin.

27. This example is drawn from Frank Jackson, "Epiphenomenal Qualia."

28. One good way of getting a sense of what's going on in consciousness research is to visit the Web site of the Association for the Scientific Study of Consciousness (ASSC) and download the program of a recent annual conference. The programs have a list of lectures, symposia, and contributed papers with informative abstracts.

29. In a recent book, *Mind and Consciousness: 5 Questions*, ed. Patrick Grim, twenty prominent philosophers of mind are asked the question "Is a science of consciousness possible?" Several philosophers give an unqualified "yes, of course" answer; almost all give affirmative answers, and no one a flat-out no answer. However, many of the respondents may have had in mind access consciousness, not phenomenal consciousness.

30. Frank Jackson, "Epiphenomenal Qualia." The quoted paragraphs are from p. 765 of *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

31. On recognitional concepts, see Brian Loar, "Phenomenal States," in *The Nature of Consciousness*, ed. Block, Flanagan, and Güzeldere, pp. 600ff.

32. See Lawrence Nemirow, "So *This* Is What It's Like: A Defense of the Ability Hypothesis"; David Lewis, "What Experience Teaches."

33. How about the proposition "Tomatoes don't look like lemons"? Is this a piece of new, demonstrative-free information that Mary can gain on her release? No, this is something Mary could know in her black-and-white room.

She knew all about the wavelengths of reflected light from tomatoes and lemons and how these wavelengths correspond to the different visual looks of objects. She only lacked knowledge of what it is like to visually experience these looks and how they differ from each other.

34. Jackson himself has renounced the knowledge argument. He now embraces a more physicalist-friendly stance; see his "The Knowledge Argument, Diaphanousness, Representationalism."

35. But what of the causal powers of pain as such—that is, as a mental kind? Strictly speaking, causation is a relation between instances of properties—that is, individual events and states—not between properties. This means that once we have vindicated the causal efficacy of each instance of a mental property, there is no further issue of vindicating the causal efficacy of the property "as such." Because mental kinds and properties are subject to multiple realization, we have to expect mental kinds to be highly causally heterogeneous, and we cannot identify the causal powers of a mental property or kind with those of any single physical property or kind. For more details, see Jaegwon Kim, "Reduction and Reductive Explanation: Is One Possible Without the Other?"

36. See David J. Chalmers, *The Conscious Mind*; Jaegwon Kim, *Physicalism, or Something Near Enough*.

37. In fact, the question of metaphysical possibility may well be irrelevant here. Since the issue is the definability of mental terms, it is a conceptual issue, and the conceivability of spectrum inversion suffices to show the indefinability of color qualia in behavioral-functional terms.

38. This was discussed earlier, in connection with qualia supervenience.

39. Note that we have only produced reasons for being unpersuaded by these arguments for identity reduction of qualia; we have not shown that identity reduction cannot work. (See note 40 on qualia and multiple realization.) This opens up an intriguing possibility: Intentional-cognitive states are reduced by functional reduction and qualia are reduced by identity reduction. This would cover all of mentality, and we would be home free! However, we must set aside further discussion of this strategy.

40. Doesn't the multiple realization argument actually defeat the identity reduction of qualia? Although Hilary Putnam used the case of pain to formulate his multiple realization argument (chapter 4), the argument works best for intentional-cognitive states. It is not implausible to link qualia closely to their neural-biological bases and deny their multiple realizability. See Christopher Hill, *Consciousness*, pp. 30–31.

41. Jerry A. Fodor, "Making Mind Matter More," in Fodor, *A Theory of Content and Other Essays*, p. 156.

References

Alanen, Lilli. *Descartes's Concept of Mind* (Cambridge, MA: Harvard University Press, 2003).

Alexander, Samuel. *Space, Time, and Deity,* 2 vols. (London: Macmillan, 1920).

Allen, Colin. "It Isn't What You Think: A New Idea About Intentional Causation," *Noûs* 29 (1995): 115–126.

Alter, Torin, and Robert J. Howell. *A Dialogue on Consciousness* (Oxford: Oxford University Press, 2009).

Alter, Torin, and Sven Walter, eds. *Phenomenal Concepts and Phenomenal Knowledge* (Oxford: Oxford University Press, 2007).

Antony, Louise. "Anomalous Monism and the Problem of Explanatory Force," *Philosophical Review* 98 (1989): 153–188.

_____. "Everybody Has Got It: A Defense of Non-Reductive Materialism," in *Contemporary Debates in Philosophy of Mind*, ed. Brian P. McLaughlin and Jonathan Cohen.

Armstrong, David. "The Nature of Mind," in *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block.

Armstrong, David M., and Norman Malcolm. *Consciousness and Causality* (Oxford: Blackwell, 1984).

Baars, Bernard J. *In the Theater of Consciousness: The Workspace of the Mind* (New York: Oxford University Press, 1997).

Bailey, Andrew M., Joshua Rasmussen, and Luke Van Horn, "No Pairing Problem," *Philosophical Studies*, forthcoming.

Baker, Lynne Rudder. *Explaining Attitudes* (Cambridge: Cambridge University Press, 1995).

_____. "Has Content Been Naturalized?" in *Meaning in Mind,* ed. Barry Loewer and Georges Rey.

Balog, Katalin. "Phenomenal Concepts," in *The Oxford Handbook of Philosophy of Mind*, ed. Brian McLaughlin et al.

Beakley, Brian, and Peter Ludlow, eds. *The Philosophy of Mind*, 2nd ed. (Cambridge, MA: MIT Press, 2006).

Bechtel, William, and Jennifer Mundale. "Multiple Realizability Revisited: Linking Cognitive and Neural States," *Philosophy of Science* 66 (1999): 175–207.

Bennett, Karen. "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It," *Noûs* 37 (2003): 471–497.

_____. "Mental Causation," *Philosophical Compass* 2 (2007): 316–337.

_____. "Exclusion Again," in *Being Reduced*, ed. Jakob Hohwy and Jesper Kallestrup.

Block, Ned. "Troubles with Functionalism," *Minnesota Studies in the Philosophy of Science*, vol. 9 (1978): 261–325. Reprinted in *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block; and Block, *Consciousness, Function, and Representation*.

_____. "What Is Functionalism?" in *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block. Reprinted in Block, *Consciousness, Function, and Representation*; *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

_____. "Psychologism and Behaviorism," *Philosophical Review* 90 (1981): 5–43.

_____. "Can the Mind Change the World?" in *Meaning and Method*, ed. George Boolos (Cambridge: Cambridge University Press, 1990).

_____. "Inverted Earth," *Philosophical Perspectives* 4 (1990): 51–79.

_____. "On a Confusion About a Function of Consciousness," *Behavioral and Brain Sciences* 18 (1995): 1–41. Reprinted in *The Nature of Consciousness,* ed. Ned Block, Owen Flanagan, and Güven Güzeldere; and in Block, *Consciousness, Function, and Representation*.

_____. "The Mind as Software in the Brain," in *An Invitation to Cognitive Science*, ed. Daniel N. Osherson (Cambridge, MA: MIT Press, 1995). Reprinted in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

_____. "Anti-Reductionism Slaps Back," *Philosophical Perspectives* 11 (1997): 107–132.

_____. "The Harder Problem of Consciousness," *Journal of Philosophy* 94 (2002): 1–35. A longer version is reprinted in Block, *Consciousness, Function, and Representation*.

_____. "Mental Paint," in *Reflections and Replies*, ed. Martin Hahn and Bjorn Ramberg (Cambridge, MA: MIT Press, 2003). Reprinted in Block, *Consciousness, Function, and Representation.*

_____. "Concepts of Consciousness," in Block, *Consciousness, Function, and Representation*.

_____. *Consciousness, Function, and Representation* (Cambridge, MA: MIT Press, 2007).

_____, ed. *Readings in Philosophy of Psychology*, vol. 1 (Cambridge, MA: Harvard University Press, 1980).

Block, Ned, Owen Flanagan, and Güven Güzeldere, eds. *The Nature of Consciousness: Philosophical and Scientific Essays* (Cambridge, MA: MIT Press, 1999).

Block, Ned, and Robert Stalnaker. "Conceptual Analysis, Dualism, and the Explanatory Gap," *Philosophical Review* 108 (1999): 1–46. Reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.

Boghossian, Paul. "Content and Self-Knowledge," *Philosophical Topics* 17 (1989): 5–26.

_____. "Naturalizing Content," in *Meaning in Mind,* ed. Barry Loewer and Georges Rey.

Boolos, George S., John Burgess, and Richard C. Jeffrey. *Computability and Logic*, 4th ed. (Cambridge: Cambridge University Press, 2002).

Borchert, Donald, ed. *The Macmillan Encyclopedia of Philosophy*, 2nd ed. (New York: Macmillan, 2005).

Brentano, Franz. *Psychology from an Empirical Standpoint*, trans. Antos C. Rancurello, D. B. Terrell, and Linda L. McAlister (New York: Humanities Press, 1973).

Burge, Tyler. "Individualism and the Mental," *Midwest Studies in Philosophy* 4 (1979): 73–121. Reprinted in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil. An excerpted version appears in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.

_____. "Individualism and Self-Knowledge," *Journal of Philosophy* 85 (1988): 654–655. Reprinted in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

Byrne, Alex. "Intentionalism Defended," *Philosophical Review* 110 (2001): 199–240.

Carnap, Rudolf. "Psychology in Physical Language," in *Logical Positivism*, ed. A. J. Ayer (New York: Free Press, 1959). First published in 1932 in German.

Carruthers, Peter. *Consciousness: Essays from a Higher-Order Perspective* (Oxford: Clarendon Press, 2005).

_____. "Higher-Order Theories of Consciousness," *Stanford Encyclopedia of Philosophy*, 2007 (http://plato.stanford.edu).

Carruthers, Peter, and Venedicte Veillet. "The Phenomenal Concept Strategy," *Journal of Consciousness Studies* 14 (2007): 212–236.

Chalmers, David J. *The Conscious Mind* (New York: Oxford University Press, 1996).

_____, ed. *Philosophy of Mind: Classical and Contemporary Readings* (Oxford: Oxford University Press, 2002).

Chisholm, Roderick M. *Perceiving* (Ithaca, NY: Cornell University Press, 1957).

_____. *The First Person* (Minneapolis: University of Minnesota Press, 1981).

Chomsky, Noam. Review of B. F. Skinner, *Verbal Behavior*. *Language* 35 (1959): 26–58.

Churchland, Patricia S. "Can Neurobiology Teach Us Anything About Consciousness?" in *The Nature of Consciousness*, ed. Ned Block, Owen Flanagan, and Güven Gülzedere. First published in 1994.

Churchland, Paul M. "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy* 78 (1981): 67–90. Reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers; *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

Clark, Andy. *Mindware: An Introduction to the Philosophy of Cognitive Science* (New York and Oxford: Oxford University Press, 2001).

Cottingham, John, Robert Stoothoff, and Dugald Murdoch, eds. *The Philosophical Writings of Descartes*, 3 vols. (Cambridge: Cambridge University Press, 1985).

Craig, Edward, ed. *The Routledge Encyclopedia of Philosophy* (London: Routledge, 1998).

Crane, Tim. "The Causal Efficacy of Content: A Functionalist Theory," in *Human Action, Deliberation, and Causation*, ed. Jan Bransen and Stefaan E. Cuypers (Dordrecht: Kluwer, 1998).

_____. "Mental Substances," in *Minds and Persons*, ed. Anthony O'Hear (Cambridge: Cambridge University Press, 2003).

Crick, Francis. *The Astonishing Hypothesis* (New York: Scribner, 1995).

Crumley II, Jack S., ed. *Problems in Mind* (Mountain View, CA: Mayfield, 2000).

Cummins, Denise Dellarosa, and Robert Cummins, eds. *Minds, Brains, and Computers: An Anthology* (Oxford: Blackwell, 2000).

Cummins, Robert. *Meaning and Mental Representation* (Cambridge, MA: MIT Press, 1989).

Davidson, Donald. "Actions, Reasons, and Causes" (1963), reprinted in *Essays on Actions and Events*, ed. Donald Davidson (New York: Oxford University Press, 1980).

_____. "The Individuation of Events" (1969), reprinted in *Essays on Actions and Events*, ed. Donald Davidson.

_____. "Mental Events" (1970), reprinted in Davidson, *Essays on Actions and Events*; in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers; *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

_____. "Radical Interpretation" (1973), reprinted in Davidson, *Inquiries into Truth and Interpretation*; *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

_____. "Belief and the Basis of Meaning" (1974), reprinted in Davidson, *Inquiries into Truth and Interpretation*.

_____. "Thought and Talk" (1974), reprinted in Davidson, *Inquiries into Truth and Interpretation*; *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

_____. *Essays on Actions and Events* (New York: Oxford University Press, 1980).

_____. "Rational Animals" (1982), reprinted in Davidson, *Subjective, Intersubjective, Objective*.

_____. *Inquiries into Truth and Interpretation* (New York: Oxford University Press, 1984).

_____. "Knowing One's Own Mind" (1987), reprinted in Davidson, *Subjective, Intersubjective, Objective*.
_____. "Three Varieties of Knowledge" (1991), reprinted in Davidson, *Subjective, Intersubjective Objective*.
_____. "Thinking Causes," in *Mental Causation*, ed. John Heil and Alfred Mele.
_____. *Subjective, Intersubjective, Objective* (Oxford: Clarendon, 2001).
Davis, Martin. *Computability and Unsolvability* (New York: McGraw-Hill, 1958).
Dennett, Daniel C. *Brainstorms* (Montgomery, VT: Bradford Books, 1978).
_____. "Intentional Systems," reprinted in Dennett, *Brainstorms*.
_____. "True Believers," in Daniel C. Dennett, *Intentional Stance* (Cambridge, MA: MIT Press, 1987). Reprinted in *The Nature of Mind*, ed. David Rosenthal; *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.
_____. "Quining Qualia," in *Consciousness in Contemporary Science*, ed. A. J. Marcel and E. Bisiach. Reprinted in *The Nature of Consciousness*, ed. Ned Block, Owen Flanagan, and Güven Güzeldere; *Readings in Philosophy and Cognitive Science*, ed. Alvin Goldman.
_____. *Consciousness Explained* (Boston: Little, Brown, 1991).
Descartes, René. *Meditations on First Philosophy*, in *The Philosophical Writings of Descartes,* vol. 2, ed. John Cottingham, Robert Stoothoff, and Dugald Murdoch.
_____. *The Passions of the Soul*, book 1, in *The Philosophical Writings of Descartes,* vol. 1, ed. John Cottingham, Robert Stoothoff, and Dugald Murdoch.
_____. "Author's Replies to the Second Set of Objections," in *The Philosophical Writings of Descartes*, vol. 2, ed. John Cottingham, Robert Stoothoff, and Dugald Murdoch.
_____. "Author's Replies to the Fourth Set of Objections," in *The Philosophical Writings of Descartes,* vol. 2, ed. John Cottingham, Robert Stoothoff, and Dugald Murdoch.
Dretske, Fred. *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981).
_____. "Misrepresentation," in *Belief*, ed. Radu Bogdan (Oxford: Oxford University Press, 1986); reprinted in *Readings in Philosophy and Cognitive Science*, ed. Alvin Goldman.
_____. *Explaining Behavior* (Cambridge, MA: MIT Press, 1988).
_____. *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995).
_____. "Minds, Machines, and Money: What Really Explains Behavior," in *Human Action, Deliberation, and Causation*, ed. Jan Bransen and Stefaan E. Cuypers (Dordrecht: Kluwer, 1998).
Egan, Frances. "Must Psychology Be Individualistic?" *Philosophical Review* 100 (1991): 179–203.
Enç, Berent. "Redundancy, Degeneracy, and Deviance in Action," *Philosophical Studies* 48 (1985): 353–374.

Feigl, Herbert. *The "Mental" and the "Physical": The Essay and a Postscript* (Minneapolis: University of Minnesota Press, 1967). First published in 1958; excerpted in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.

Flanagan, Owen. *Consciousness Reconsidered* (Cambridge, MA: MIT Press, 1992).

Fodor, Jerry A. "Special Sciences, or the Disunity of Science as a Working Hypothesis," *Synthese* 28 (1974): 97–115. Reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.

_____. *Psychosemantics* (Cambridge, MA: MIT Press, 1987).

_____. *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990).

_____. "Making Mind Matter More," in Fodor, *A Theory of Content and Other Essays*.

_____. "A Modal Argument for Narrow Content," *Journal of Philosophy* 88 (1991): 5–26.

_____. "Special Sciences: Still Autonomous After All These Years," reprinted in Fodor, *A Critical Condition* (Cambridge, MA: MIT Press, 2000). First published in 1997.

Foster, John. *The Case for Idealism* (London: Routledge, 1982).

_____. *The Immaterial Self* (London: Routledge, 1991).

_____. "A Defense of Dualism," in *The Case for Dualism*, ed. John R. Smythies and John Beloff (Charlottesville: University Press of Virginia, 1989). Reprinted in *Problems in Mind*, ed. Jack S. Crumley II.

_____. "A Brief Defense of the Cartesian View," in *Soul, Body, and Survival*, ed. Kevin Corcoran (Ithaca, NY: Cornell University Press, 2001).

Garber, Daniel. "Understanding Interaction: What Descartes Should Have Told Elisabeth," in Garber, *Descartes Embodied*.

_____. *Descartes Embodied* (Cambridge: Cambridge University Press, 2001).

Gendler, Tamar Szabo, and John Hawthorne, eds. *Conceivability and Possibility* (Oxford: Oxford University Press, 2002).

Gibbons, John. "Mental Causation Without Downward Causation," *Philosophical Review* 115 (2006): 79–103.

Gillett, Carl, and Barry Loewer, eds. *Physicalism and Its Discontents* (Cambridge: Cambridge University Press, 2001)

Ginet, Carl. *On Action* (Cambridge: Cambridge University Press, 1990).

Goldman, Alvin I. "Interpretation Psychologized," in Goldman, *Liaisons* (Cambridge, MA: MIT Press, 1992). First published in 1989.

_____. "Consciousness, Folk Psychology, and Cognitive Science," *Consciousness and Cognition* 2 (1993): 364–382. Reprinted in *The Nature of Consciousness*, ed. Ned Block, Owen Flanagan, and Güven Gülzedere.

_____. *Simulating Minds* (Oxford: Oxford University Press, 2006).

_____, ed. *Readings in Philosophy and Cognitive Science* (Cambridge, MA: MIT Press, 1993).

Gopnik, Alison. "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality," *Behavioral and Brain Sciences* 16 (1993): 1–14. Reprinted in *Readings in Philosophy and Cognitive Science*, ed. Alvin I. Goldman.

Gordon, Robert M. "Folk Psychology as Simulation," *Mind and Language* 1 (1986): 159–171.

Grim, Patrick, ed. *Mind and Consciousness: 5 Questions* (Automatic Press, 2009).

Hardin, C. L. *Color for Philosophers* (Indianapolis: Hackett, 1988).

Harman, Gilbert. "The Inference to the Best Explanation," *Philosophical Review* 74 (1966): 88–95.

_____. "The Intrinsic Quality of Experience," *Philosophical Perspectives* 4 (1990): 31–52. Reprinted in *The Nature of Consciousness,* ed. Ned Block, Owen Flanagan, and Güven Güzeldere.

Harnish, Robert M. *Minds, Brains, Computers: An Historical Introduction to the Foundations of Cognitive Science* (Oxford: Blackwell, 2002).

Hart, W. D. *The Engines of the Soul* (Cambridge: Cambridge University Press, 1988).

Hasker, William. *The Emergent Self* (Ithaca, NY: Cornell University Press, 1999).

Heil, John. *The Nature of True Minds* (Cambridge: Cambridge University Press, 1992).

_____, ed. *Philosophy of Mind: A Guide and Anthology* (Oxford: Oxford University Press, 2004).

Heil, John, and Alfred Mele, eds. *Mental Causation* (Oxford: Clarendon Press, 1993).

Hempel, Carl G. "The Logical Analysis of Psychology" (1935), in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

_____. *Philosophy of Natural Science* (Englewood Cliffs, NJ: Prentice-Hall, 1966).

Hill, Christopher S. *Sensations: A Defense of Type Materialism* (Cambridge: Cambridge University Press, 1991).

_____. *Consciousness* (Cambridge: Cambridge University Press, 2009).

Hohwy, Jakob, and Jesper Kallestrup, eds. *Being Reduced* (Oxford: Oxford University Press, 2008).

Horgan, Terence. "Supervenient Qualia," *Philosophical Review* 96 (1987): 491–520.

_____. "Mental Quausation," *Philosophical Perspectives* 3 (1989): 47–76.

_____. "From Supervenience to Superdupervenience: Meeting the Demands of a Material World," *Mind* 102 (1993): 555–586.

Huxley, Thomas H. *Lessons in Elementary Physiology* (London: Macmillan, 1885).

_____. "On the Hypothesis That Animals Are Automata, and Its History," excerpted in *Philosophy of Mind: Classical and Contemporary Readings*, ed.

David J. Chalmers. A full version appears in *Methods and Results: Essays by Thomas H. Huxley* (New York: D. Appleton, 1901).

Jackson, Frank. "Finding the Mind in the Natural World" (1994), reprinted in *The Nature of Consciousness,* ed. Ned Block, Owen Flanagan, and Güven Güzeldere.

_____. "Epiphenomenal Qualia," *Philosophical Quarterly* 32 (1982): 127–138. Reprinted in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

_____. "The Knowledge Argument, Diaphanousness, Representationalism," in *Phenomenal Concepts and Phenomenal Knowledge*, ed. Torin Alter and Sven Walter.

Jacob, Pierre. *What Minds Can Do* (Cambridge: Cambridge University Press, 1997).

James, William. *The Principles of Psychology* (1890; Cambridge, MA: Harvard University Press, 1981).

Jolley, Nicholas. *Locke: His Philosophical Thought* (Oxford: Oxford University Press, 1999).

Kim, Jaegwon. "Events as Property Exemplifications" (1976), reprinted in Kim, *Supervenience and Mind.*

_____. "Psychophysical Laws" (1985), reprinted in Kim, *Supervenience and Mind.*

_____. "The Myth of Nonreductive Materialism," reprinted in Kim, *Supervenience and Mind.*

_____. "Multiple Realization and the Metaphysics of Reduction" (1992), reprinted in Kim, *Supervenience and Mind*; in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers; *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

_____. *Supervenience and Mind* (Cambridge: Cambridge University Press, 1993).

_____. *Mind in a Physical World* (Cambridge, MA: MIT Press, 1998).

_____. *Physicalism, or Something Near Enough* (Princeton, NJ: Princeton University Press, 2005).

_____. "Reduction and Reductive Explanation: Is One Possible Without the Other?" in *Being Reduced*, ed. Jakob Hohwy and Jesper Kallestrup. Reprinted in Kim, *Essays in the Metaphysics of Mind.*

_____. "Why There Are No Laws in the Special Sciences: Three Arguments," in Kim, *Essays in the Metaphysics of Mind.*

_____. *Essays in the Metaphysics of Mind* (Oxford: Oxford University Press, 2010).

_____. "The Very Idea of Token Physicalism," in *New Perspectives on Type Physicalism*, ed. Simone Gozzano and Christopher Hill (Cambridge: Cambridge University Press, forthcoming).

Kind, Amy. "What's So Transparent About Transparency?" *Philosophical Studies* 115 (2003): 225–244.

_____. "Restrictions on Representationalism," *Philosophical Studies* 134 (2007): 405–427.

Koons, Robert C., and George Bealer. *The Waning of Materialism* (Oxford: Oxford University Press, 2010).

Kripke, Saul. *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980).

Lashley, Karl. *Brain Mechanisms and Intelligence* (New York: Hafner, 1963).

Latham, Noa. "Substance Physicalism," in *Physicalism and Its Discontents*, ed. Carl Gillett and Barry Loewer.

Leibniz, Gottfried. *Monadology*, 1714. Various editions and translations.

LePore, Ernest, and Barry Loewer. "Mind Matters," *Journal of Philosophy* 84 (1987): 630–642.

Levin, Janet. "Could Love Be Like a Heatwave?" *Philosophical Studies* 49 (1986): 245–261.

Levine, Joseph. "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64 (1983): 354–361. Reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers; *Philosophy of Mind: A Guide and Anthology,* ed. John Heil.

_____. "On Leaving Out What It's Like," in *Consciousness,* ed. Martin Davies and Glyn W. Humphreys (Oxford: Blackwell, 1993).

_____. *Purple Haze* (Oxford: Oxford University Press, 2000).

Lewis, David. "An Argument for the Identity Theory," *Journal of Philosophy* 63 (1966): 17–25. Reprinted in Lewis, *Philosophical Papers*, vol. 1.

_____. "How to Define Theoretical Terms" (1970), reprinted in Lewis, *Philosophical Papers*, vol. 1.

_____. *Counterfactuals* (Cambridge, MA: Harvard University Press, 1973).

_____. "Psychophysical and Theoretical Identifications" (1972), *Australasian Journal of Philosophy* 50 (1972): 249–258. Reprinted in Lewis, *Papers in Metaphysics and Epistemology*.

_____. "Causation" (1973), reprinted, with "Postscripts," in Lewis, *Philosophical Papers*, vol. 2.

_____. "Radical Translation," *Synthese* 27 (1974): 331–344. Reprinted in Lewis, *Philosophical Papers*, vol. 1.

_____. *Philosophical Papers*, vol. 1 (New York: Oxford University Press, 1983).

_____. "Attitudes *De Dicto* and *De Se*," *Philosophical Review* 88 (1979): 513–543. Reprinted in Lewis, *Philosophical Papers*, vol. 1.

_____. *Philosophical Papers*, vol. 2 (New York: Oxford University Press, 1986).

_____. "What Experience Teaches," *Proceedings of the Russellian Society* 13 (1988): 29–57. Reprinted in Lewis, *Papers in Metaphysics and Epistemology*; *Philosophy of Mind: Classical and Contemporary Readings*, ed. David Chalmers.

_____. *Papers in Metaphysics and Epistemology* (Cambridge: Cambridge University Press, 1999).

List, Christian, and Peter Menzies. "Nonreductive Physicalism and the Limits of the Exclusion Principle," *Journal of Philosophy* 106 (2009): 475–502.

Loar, Brian. "Phenomenal States," *Philosophical Perspectives* (1990): 81–108. Reprinted in *The Nature of Consciousness*, ed. Ned Block, Owen Flanagan, and Güven Güzeldere.

Locke, John. *An Essay Concerning Human Understanding*, ed. P. H. Nidditch (1689; New York: Oxford University Press, 1975).

Loewer, Barry, and Georges Rey, eds. *Meaning in Mind* (London: Routledge, 1991).

Lowe, E. J. "Physical Causal Closure and the Invisibility of Mental Causation," in *Physicalism and Mental Causation*, ed. Sven Walter and Heinz-Dieter Heckmann.

_____. "Non-Cartesian Substance Dualism and the Problem of Mental Causation," *Erkenntnis* 65 (2006): 5–23.

_____. "Dualism," in *The Oxford Handbook of Philosophy of Mind*, ed. Brian McLaughlin et al.

Ludlow, Peter, and Norah Martin, eds. *Externalism and Self-Knowledge* (Stanford, CA: CSLI Publications, 1998).

Ludlow, Peter, Yujin Nagasawa, and Daniel Stoljar, eds. *There's Something About Mary* (Cambridge, MA: MIT Press, 2004).

Lycan, William G. *Consciousness* (Cambridge, MA: MIT Press, 1987).

_____. *Consciousness and Experience* (Cambridge, MA: MIT Press, 1996).

Lycan, William G., and Jesse Prinz, eds. *Mind and Cognition: An Anthology*, 3rd ed. (Oxford: Blackwell, 2008).

Macdonald, Cynthia, and Graham Macdonald. "The Metaphysics of Mental Causation," *Journal of Philosophy* 103 (2006): 539–576.

Marcel, A. J., and E. Bisiach, eds. *Consciousness in Contemporary Science* (Oxford: Oxford University Press, 1988).

Marras, Ausonio. "Nonreductive Physicalism and Mental Causation," *Canadian Journal of Philosophy* 24 (1994): 465–493.

Matthews, Robert. "The Measure of Mind," *Mind* 103 (1994): 131–146.

McGinn, Colin. "Can We Solve the Mind-Body Problem?" in McGinn, *The Problem of Consciousness* (Oxford: Blackwell, 1991).

McLaughlin, Brian. "What Is Wrong with Correlational Psychosemantics?" *Synthese* 70 (1987): 271–286.

_____. "Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical," *Philosophical Perspectives* 3 (1989): 109–136.

_____. "In Defense of New Wave Materialism: A Response to Horgan and Tienson," in *Physicalism and Its Discontents*, ed. Carl Gillett and Barry Loewer.

_____. "Is Role-Functionalism Committed to Epiphenomenalism?" *Journal of Consciousness Studies* 13, no. 1–2, ed. Michael Pauen, Alexander Staudacher, and Sven Walter.

McLaughlin, Brian, Ansgar Beckermann, and Sven Walter, eds. *The Oxford Handbook of Philosophy of Mind* (Oxford: Oxford University Press, 2009).

McLaughlin, Brian, and Karen Bennett. "Supervenience," in *Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu).

McLaughlin, Brian P., and Jonathan Cohen, eds. *Contemporary Debates in Philosophy of Mind* (Oxford: Blackwell, 2007).

Melnyk, Andrew. *A Physicalist Manifesto* (Cambridge: Cambridge University Press, 2003).

_____. "Can Physicalism Be Non-Reductive?" *Philosophy Compass* 3, no. 6 (2008): 1281–1296.

Mendola, Joseph. *Anti-Externalism* (Oxford: Oxford University Press, 2008).

Millikan, Ruth G. *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984).

_____. "Biosemantics," *Journal of Philosophy* 86 (1989): 281–297. Reprinted in *Problems in Mind*, ed. Jack S. Crumley II; and in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.

Nagel, Thomas. "What Is It Like to Be a Bat?" *Philosophical Review* 83 (1974): 435–450. Reprinted in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil; *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.

_____. "Subjective and Objective," in Thomas Nagel, *Mortal Questions* (Cambridge: Cambridge University Press, 1979).

_____. *The View from Nowhere* (Oxford: Oxford University Press, 1986).

Neander, Karen. "Teleological Theories of Mental Content," in *Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu).

Nemirow, Lawrence. "So *This* Is What It's Like: A Defense of the Ability Hypothesis," in *Phenomenal Concepts and Phenomenal Knowledge*, ed. Torin Alter and Sven Walter.

Ney, Alyssa. "Defining Physicalism," *Philosophy Compass* 3 (2008): 1033–1048.

Nida-Rümelin, Martine. "Pseudo-Normal Vision: An Actual Case of Qualia Inversion?" *Philosophical Studies* 82 (1996): 145–157. Reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.

Nisbett, Richard E., and Timothy DeCamp Wilson. "Telling More Than We Can Know," *Psychological Review* 84 (1977): 231–259.

Nuccetelli, Susan, ed. *New Essays on Semantic Externalism and Self-Knowledge* (Cambridge, MA: MIT Press, 2003).

O'Connor, Timothy, and David Robb, eds. *Philosophy of Mind: Contemporary Readings* (London: Routledge, 2003).

Olson, Eric T. *The Human Animal: Personal Identity Without Psychology* (Oxford: Oxford University Press, 1997).

Papineau, David. "The Rise of Physicalism," in *Physicalism and Its Discontents*, ed. Carl Gillett and Barry Loewer.

_____. *Thinking About Consciousness* (Oxford: Oxford University Press, 2002).

_____. "The Causal Closure of the Physical and Naturalism," in *The Oxford Handbook of Philosophy of Mind*, ed. Brian McLaughlin et al.

Pauen, Michael, Alexander Staudacher, and Sven Walter, eds. *Consciousness Studies: Special Issue on Epiphenomenalism*, vol. 13, no. 1–2 (2006).

Pavlov, Ivan. *Experimental Psychology and Other Essays* (New York: Philosophical Library, 1957), p. 148.

Perry, John. *Knowledge, Possibility, and Consciousness* (Cambridge, MA: MIT Press, 2001).

Plantinga, Alvin. "Against Materialism," *Faith and Philosophy* 23 (2006): 3–32.

Poland, Jeffrey. *Physicalism: The Philosophical Foundation* (Oxford: Clarendon Press, 1994).

Polger, Thomas W. *Natural Minds* (Cambridge, MA: MIT Press, 2004).

Proust, Marcel. *Remembrance of Things Past*, vol. 1, trans. C. K. Scott Moncrieff and Terence Kilmartin (New York: Vintage, 1982).

Putnam, Hilary. "Brains and Behavior" (1965), reprinted in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil; and in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David. J. Chalmers.

_____. "Psychological Predicates," in *Art, Mind, and Religion*, ed. W. H. Capitan and D. D. Merrill (Pittsburgh: University of Pittsburgh Press, 1967). Retitled as "The Nature of Mental States" and reprinted in Putnam, *Mind, Language, and Reality: Philosophical Papers*, vol. 2. Also in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil; *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.

_____. "Robots: Machines or Artificially Created Life?" (1964), in *Mind, Language, and Reality: Philosophical Papers*, vol. 2.

_____. "The Meaning of 'Meaning'" (1975), reprinted in Putnam, *Mind, Language, and Reality: Philosophical Papers*, vol. 2. An excerpted version appears in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.

_____. *Mind, Language, and Reality: Philosophical Papers*, vol. 2, 2nd ed. (Cambridge: Cambridge University Press, 1979).

_____. *Representation and Reality* (Cambridge, MA: MIT Press, 1988).

Quine, W. V. *Word and Object* (Cambridge and New York: Technology Press of MIT and John Wiley & Sons, 1960).

Rey, Georges. "A Question about Consciousness," reprinted in *The Nature of Consciousness*, ed. Ned Block, Owen Flanagan, and Güven Güzeldere. First published in 1988.

Rimbaud, Arthur. "Voyelles," in *Arthur Rimbaud: Complete Works*, trans. Paul Schmidt (New York: Harper & Row, 1976).

Rosenthal, David M. "The Independence of Consciousness and Sensory Quality," *Philosophical Issues* 1 (1991): 15–36.

_____. "Explaining Consciousness," in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.

_____. *Consciousness and Mind* (Oxford: Oxford University Press, 2006).

Ross, Don, and David Spurrett. "What to Say to a Skeptical Metaphysician: A Defense Manual for Cognitive and Behavioral Scientists," *Behavioral and Brain Sciences* 27 (2004): 603–647.

Rowlands, Mark. "Consciousness and Higher-Order Thoughts," *Mind and Language* 16 (2001): 290–310.

Rozemond, Marleen. *Descartes's Dualism* (Cambridge, MA: Harvard University Press, 1998).

Ryle, Gilbert. *The Concept of Mind* (New York: Barnes and Noble, 1949).

Searle, John. "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3 (1980): 417–424. Reprinted in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil; *Philosophy of Mind: Contemporary Readings*, ed. Timothy O'Connor and David Robb.

_____. *Intentionality* (Cambridge: Cambridge University Press, 1983).

_____. *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992).

Segal, Gabriel M. A. *A Slim Book About Narrow Content* (Cambridge, MA: MIT Press, 2000).

Shaffer, Jerome. "Mental Events and the Brain," *Journal of Philosophy* 60 (1963): 160–166. Reprinted in *The Nature of Mind*, ed. David M. Rosenthal.

Shapiro, Lawrence. *The Mind Incarnate* (Cambridge, MA: MIT Press, 2004).

Shoemaker, Sydney. "The Inverted Spectrum," *Journal of Philosophy* 79 (1982): 357–382. Reprinted in Shoemaker, *Identity, Cause, and Mind*.

_____. "Some Varieties of Functionalism," in Shoemaker, *Identity, Cause, and Mind*.

_____. "Absent Qualia Are Impossible—A Reply to Block," in Shoemaker, *Identity, Cause, and Mind*.

_____. *Identity, Cause, and Mind* (Cambridge: Cambridge University Press, 1984).

_____. *Physical Realization* (Oxford: Oxford University Press, 2008).

Siewert, Charles. "Is Experience Transparent?" *Philosophical Studies* 117 (2004): 15–41.

Skinner, B. F. "Selections from *Science and Human Behavior*" (1953), reprinted in *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block.

_____. *Science and Human Behavior* (New York: Macmillan, 1953).

_____. *About Behaviorism* (New York: Alfred A. Knopf, 1974).

Smart, J. J. C. "Sensations and Brain Processes," *Philosophical Review* 68 (1959): 141–156. Reprinted in *The Nature of Mind*, ed. David M. Rosenthal; *Philosophy of Mind: A Guide and Anthology*, ed. John Heil; *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers.

Smith, Michael. "The Possibility of Philosophy of Action," in *Human Action, Deliberation, and Causation*, ed. Jan Bransen and Stefaan E. Cuypers.

Sosa, Ernest. "Mind-Body Interaction and Supervenient Causation," *Midwest Studies in Philosophy* 9 (1984): 271–281.

_____. "Between Internalism and Externalism," *Philosophical Issues* 1 (1991): 179–195.

Stalnaker, Robert. *Inquiry* (Cambridge, MA: MIT Press, 1984).

Stampe, Dennis. "Toward a Causal Theory of Linguistic Representation," *Midwest Studies in Philosophy* 2 (1977): 42–63.

Stanford Online Encyclopedia of Philosophy (http://plato.stanford.edu).

Stich, Stephen P. *From Folk Psychology to Cognitive Science: The Case Against Belief* (Cambridge, MA: MIT Press, 1983).

Stoljar, Daniel. *Physicalism* (London and New York: Routledge, 2010).

Stoutland, Frederick. "Oblique Causation and Reasons for Action," *Synthese* 43 (1980): 351–367.

_____. "Real Reasons," in *Human Action, Deliberation, and Causation*, ed. Jan Bransen and Stefaan E. Cuypers.

Strawson, Galen. "Real Intentionality 3: Why Intentionality Entails Consciousness," in Strawson, *Real Materialism and Other Essays* (Oxford: Oxford University Press, 2008).

Stubenberg, Leopold. *Consciousness and Qualia* (Amsterdam: John Benjamins Publishing Co., 1998).

Swinburne, Richard. *The Evolution of the Soul* (Oxford: Clarendon, 1986).

Turing, Alan M. "Computing Machinery and Intelligence," *Mind* 59 (1950): 433–460. Reprinted in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil.

Tye, Michael. "Qualia, Content, and the Inverted Spectrum," *Noûs* 28 (1994): 159–183.

_____. *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995).

Van Fraassen, Bas. *The Scientific Image* (Oxford: Clarendon, 1980).

_____. *Laws and Symmetry* (Oxford: Oxford University Press, 1989).

Van Gulick, Robert. "Consciousness," *Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu).

Velmans, Max, and Susan Schneider, eds. *The Blackwell Companion to Consciousness* (Oxford: Blackwell, 2007).

Von Eckardt, Barbara. *What Is Cognitive Science?* (Cambridge, MA: MIT Press, 1992).

Walter, Sven, and Heinz-Dieter Heckmann, eds. *Physicalism and Mental Causation: The Metaphysics of Mind and Action* (Charlottesville, VA: Imprint Academic, 2003).

Watson, J. B. "Psychology as the Behaviorist Views It," *Psychological Review* 20 (1913): 158–177.

Weiskrantz, Lawrence. *Blindsight* (Oxford: Oxford University Press, 1986).

Witmer, Gene. "Multiple Realizability and Psychological Law: Evaluating Kim's Challenge," in *Physicalism and Mental Causation*, ed. Sven Walter and Heinz-Dieter Heckmann.

Wittgenstein, Ludwig. *Philosophical Investigations*, trans. G. E. M. Anscombe (Oxford: Blackwell, 1953).

Wright, Crispin, Barry C. Smith, and Cynthia Macdonald, eds. *Knowing Our Own Minds* (Oxford: Clarendon Press, 1998).

Wright, Edmond, ed. *The Case for Qualia* (Cambridge, MA: MIT Press, 2008).

Yablo, Stephen. "Mental Causation," *Philosophical Review* 101 (1992): 245–280. Reprinted in Yablo, *Thoughts*.

_____. "Wide Causation." *Philosophical Perspectives* 11 (1997): 251–281. Reprinted in Yablo, *Thoughts*.

_____. *Thoughts* (New York: Oxford University Press, 2009).

Zimmerman, Dean. "Material People," in *The Oxford Handbook of Metaphysics*, ed. Michael J. Loux and Dean Zimmerman (Oxford: Oxford University Press, 2005).

Index

Ability Hypothesis, 324, 325
Abstractness, of psychological properties, 132
Access consciousness, 272, 277–280, 319
  computational models of, 280
  as functional concept, 280
Action
  desire-belief-action principle, 73–74, 89*n*12, 190*n*8, 195–197
  involving bodily motions, 16–17, 67
  not involving bodily motions, 67
  rational, 73
Action theory, 224*n*2
Aesthetic properties, supervenience of, 10
Agency
  body and, 4–5
  mental causation and, 195–197
Alanen, Lilli, 58
Alexander, Samuel, 97, 126*n*3, 200–201, 224*n*7, 305
Allen, Colin, 258
Alphabet, of Turing machine, 139, 144–145
Alter, Torin, 127*n*19, 295, 334
Analytical behaviorism, 68
Analytical functionalism, 173
Animalism, 59*n*8
Animals
  beliefs of/intentional states of, 247–248
  consciousness of, 199–200, 287–288
  mentality of, 2
Anomalism of the mental, 203
Anomalous monism, 207, 223
  as form of epiphenomenalism, 207–209
Antirealism, about scientific theory, 149–150
Antony, Louise, 126, 223, 260*n*7
Appear
  epistemic (doxastic) sense of, 293–295
  phenomenal (sensuous) sense of, 293–295
Aristotle, 93
Armstrong, David M., 88, 134, 189, 284, 285, 299*n*36
Arthritis and tharthritis thought-experiment, 244–248
As-if/derivative intentionality, 26
Association for the Scientific Study of Consciousness, 336*n*28
Attitudes, 227
Attributes, 7
Axiom schema for identity, 107

Baars, Bernard, 278–279, 298*n*24
Babbage, Charles, 132
Baily, Andrew, 58
Baker, Lynne Rudder, 223, 258, 260*n*7

Balog, Katalin, 127*n*19
Bats, Nagel on consciousness of, 267–271
Beakley, Brian, 28
Bealer, George, 125, 334
Bechtel, William, 125
Beckermann, Ansgar, 28, 127*n*19
"Beetle in the box," 64–66, 82
Behavior
  belief, meaning, and, 229–235
  brain and, 87
  defining, 66–67
  as evidence for attribution of mental states to others, 87
  evidence for qualia and, 322
  mental language and, 87
  pain, 76–78, 80–82
  relation of mental states to, 74
  verbal, 70–72
  wide content and, 255–256
Behavioral dispositions, mental states as, 136–137
Behavior causation, 222–223
Behaviorism, 61–90
  behavioral translation of "Paul has a toothache," 70–71
  behavior defined, 66–67
  difficulties with behavioral definitions, 71–76
  functionalism and, 136–138
  logical/analytical, 68–69
  methodological, 83, 84–85
  ontological, 78–80
  pain and pain behavior, 76–78, 80–82
  philosophical/analytical, 62, 68
  in psychology, 82–86
  radical, 82
  why behavior matters to mind, 86–87
Belief-desire-action principle, 73–74, 89*n*12, 190*n*8, 195–197
Belief
  behavior, meaning, and, 229–235
  behaviorist definition of, 72
  conscious, 275–276
  content of, 235–239
  defining, 3
  desire-belief-action principle, 73–74, 89*n*12, 190*n*8, 195–197
  dispositional, 275
  function of, 330–331
  individuated by content, 244
  motivation to act and, 88–89*n*12
  narrow content, 242, 246, 251–254
  observational, 237–238
  phenomenal feel of, 275
  physical effects of, 4–5
  radical interpretation and, 229
  rationality principle and, 206
  speech practices and content of, 245–248
  verbal behavior and, 71–72
  wide content of, 242, 244, 245–246, 248–251
  See also Intentional states
Belief tokens, 253
Belief type, 253
Bennett, Karen, 28*n*1, 223–224
Biological naturalism, 164
Biological properties, supervenience on physicochemical properties, 10–11
Black, Max, 127*n*17
Blindsight, 279–280
Block, Ned, 29*n*14
  on access consciousness, 278, 279
  on causal powers of functional properties, 189
  on consciousness, 296
  on correlations, 109
  criticism of functionalism, 189
  functional-state identity theory, 191*n*14
  on phenomenal consciousness, 273
  psychological laws, 189
  on psychoneural correlations, 314

publications, 126, 126*n*12, 165, 167*n*25, 190*n*4, 191*n*9, 191*n*16, 223, 297*n*12, 297*n*13, 297*n*19, 298*n*23, 298*n*26, 299*n*45, 334, 334*n*4, 335*n*10, 336*n*21, 336*n*31
Bodily disturbance theory of pain, 299*n*47, 318–319, 336*n*24
Bodily movements, 66–67
Body
actions and, 16–17
as extended in space, 35–39
in substance dualism, 34–38
See also Mind-body problem
Boghossian, Paul, 259, 260*n*7
Boolos, George S., 166*n*17
Borchert, Donald, 28
Brain
as base of mind, 87
causal power of, 164, 167*n*30
as cause of behavior, 87
as computing machine, 151
cross-wired, 181–182
gap between phenomenal consciousness and, 302–303
mind identified with, 6
See also Mind-brain correlations; Psychoneural identity theory
Brain science. *See* Neuroscience
Brentano, Franz, 23, 29*n*12
Burge, Tyler, 258, 259, 260*n*11, 260*n*15
thought-experiment, 242, 244–248
Burgess, John P., 166*n*17
Byrne, Alex, 296

Carnap, Rudolf, 87
Carruthers, Peter, 127*n*19, 296, 299*n*42
Cartesian theater, 63–66, 88*n*5
Causal argument, for psychoneural identity theory, 110–114
Causal closure of the physical domain, 112–113, 214–217, 223, 320, 112–113
causal (explanatory) closure of the neural domain, 320
Causal-correlational approach, to content, 235–239
Causal efficacy, of mental property, 327, 337*n*35
Causal-explanatory closure of the neural-physical domain, 320
Causal-explanatory efficacy of wide content, 254–256
Causal-functional kind, mental kind as, 133–134
Causal history, 221–222
Causal interactionism, 95
Causality, substance dualism and, 50–54
Causal-nomological resemblance, 186
Causal power
of brain, 164, 167*n*30
of mental properties, 184–186, 189
Causal relevance of qualia, 310
Causal roles, 131
sensory events and, 180
Causal status of consciousness, 319–321
Causal-theoretical functionalism, 169–191
cross-wired brain, 181–182
functionalism as physicalism, 177–179
functionalist properties, disjunctive properties, causal powers, 183–186
objections and difficulties, 179–186
qualia inversion, 179–181
Ramsey-Lewis method, 170–172
roles *vs.* realizers, 186–189
underlying psychology, 172–177
Causal work, 134–135
Causation
counterfactual account of, 209–213
defined, 337*n*35
mnemic, 85
nomological account of, 206, 209
See also Mental causation
Ceteris paribus clauses, 89*n*14

Ceteris paribus laws, mental causation and, 213
C-fiber stimulation
  pain and, 99–101, 105, 106, 107–110, 307–308, 326–327, 335*n*9
Chalmers, David J.
  on hard problem of consciousness, 59*n*15, 304
  publications, 28, 125, 165, 224*n*5, 333, 334, 334*n*7, 335*n*18, 335*n*19, 337*n*36
  on reductive explanation, 311
Charity principle, principle of, 225*n*11, 230–232, 259*n*3
Chinese room argument, 161–164, 165
Chisholm, Roderick M., 88, 89*n*13, 261*n*17
Chomsky, Noam, 88
Churchland, Patricia S., 267, 297*n*12
Churchland, Paul, 80, 89*n*21, 90*n*26, 190*n*7, 226*n*22, 260*n*14
Church-Turing thesis, 166–167*n*18
Clark, Andy, 28
"Cogito" argument, 35
Cognition, computationalism and, 161–162
Cognitive science
  content-carrying intentional states and, 257–258
  defense of, 189–190
  functionalism and, 129
  nonreductive physicalism and autonomy of, 124–125
  properties of, 184
  status of, 186–189
Cognitive science properties, 184
Cognitivism, 82
Collateral effects of a common cause, 94
Color, qualia inversion and, 180, 293, 299*n*46, 308
Common cause, 94
  correlated phenomena as collateral effects of, 105
Commonsense psychology
  to anchor psychological concepts, 173, 175–176
  content-carrying states and, 257
  simulation theory of, 190*n*5
  theory theory of, 190*n*5
  See also Folk psychology
Computationalism, 160–165, 221
Conceivability, possibility and, 39, 120, 335*n*15
Concepts, 116–117
  extension of, 190*n*5
Conceptual content, 285
Conditionalization, 336*n*22
Conscious, terminological issues, 271–272
Conscious beliefs, 275–276
Consciousness, 263–299
  access (*see* Access consciousness)
  animal, 199–200, 287–288
  behaviorism and, 84–86
  causal status of, 319–321
  defining a person, 42–43
  emergence of, 97
  evolution of, 322
  higher-order perception/thought and, 283–289
  human, 199–200
  identity reduction of, 313–315
  intentionality and, 27
  irreducibility of, 271
  in material things, 39–46
  mind-body problem and, 301–302
  mystery of, 266, 305–306
  Nagel on, 267–271
  neuroscience and, 317–323
  phenomenal (*see* Phenomenal consciousness)
  supervenience on physical properties, 306–311
  reduction and reductive explanation of, 311–315

representationalism about, 27, 280, 289–295
scientific study of, 265–267, 317–323
subjectivity and, 280–283
transparency of experience and consciousness representationalism, 289–295
views on, 264–267
Consciousness representationalism, 289–295, 319
Content
causal relevance of, 222
disjunctive, 238
extrinsicness of, 221–223
representational, 24–25
See also Mental content
Content-bearing states, content-carrying states, 15, 227–228, 226*n*22
cognitive science and, 257–258
commonsense psychology and, 257
See also Belief; Desire
Content externalism, 241–248
Burge's thought-experiment, 244–248
causal-explanatory efficacy of wide content, 254–256
problems for, 254–257
Putnam's thought-experiment, 242–244, 248
wide content and self-knowledge, 256–257
Content intentionality, 24–25
Content irrealism, 234
Content realism, 235
Content relativism, 234
Content sentences, 15
Correlations, 336*n*23
explaining in science, 105–106
psychoneural, 101–105, 109–110, 126*n*12, 303–305
Cottingham, John, 58
Counterfactuals, mental causation and, 209–213, 225*n*14, 225*n*17
Craig, Edward, 28
Crane, Tim, 58, 223, 258
Crick, Francis, 266–267, 297*n*11
Cross-wired brain, 181–182
Crumley, Jack S. III, 28
Cummins, Denise Dellarosa, 28
Cummins, Robert, 28, 258

Data
introspective, 83–84
purpose in science, 83
Davidson, Donald
acceptance of nonstrict laws, 226*n*19
anomalous monism, 165*n*2, 207–209, 223
principle of charity, 225*n*11, 230–232, 259*n*3
radical interpretation, 228–235
on causation, 197
on intentionality, 29–30*n*16
mental irrealist statement, 259–260*n*5
on psychophysical laws, 203–207, 224–225*n*9, 225*n*10, 225*n*12
publications, 223, 224*n*4, 259, 259*n*1, 260*n*12, 299*n*40
on truth of speaker's utterances, 229
Davis, Martin, 166*n*17
DBA. *See* Desire-belief-action principle (DBA)
Deductive-nomological explanation, 312
Defeasibility of mental-behavioral entailments, 74–78
Demonstrative concept, 117
Demonstratives, 325
Dennett, Daniel C., 335*n*16
on Cartesian theater, 88*n*5
on consciousness, 264–265
on intentionality, 30*n*16
multiple draft theory, 279
publications, 88*n*5, 89*n*25, 165, 167*n*26, 258, 259*n*4, 296*n*4, 296*n*5, 298*n*25
on qualia, 80

Dependence, between mental and physical, 12
Derivative intentionality, 26
Descartes, René, 260*n*12
  arguments for substance dualism, 35–46
  Cartesian theater, 63–66
  causal interaction in pineal gland, 95
  on consciousness, 264, 296*n*1, 296*n*2
  infallibility and transparency of mind, 20
  on having a mind, 5–6
  interactionist substance dualism, 32–35
  on knowledge of own propositional attitudes, 20
  mentality as nonspatial, 22
  mind-body problem and, 3, 5
  on immediate awareness of own feelings, thoughts, 281
  Princess Elisabeth and, 46–50, 60*n*20, 60*n*21
  on substance, 33, 58*n*2
  See also Substance dualism
Desire, physical effects of, 4–5
Desire-belief-action principle (DBA), 73–74, 89*n*12, 190*n*8, 195–197
Differential property, 37–38
Direct knowledge of own mental states, 18–19, 63, 238–239
  Cartesian theater and, 63–64
  identity theory and, 118–119
  infallibility and transparency of, 19–22
  privacy/first-person privilege, 19, 21–22
  wide content and, 256–257
  See also First-person authority
Disjunctive content, 238
Disjunctive property, 184–186
Dispositional belief, 275
Disposition, dispositional property, 71–72
  instrumentalist analysis, 136–137
  realist analysis, 136–137
Divine intervention, 96
Double-aspect theory, 96
Dretske, Fred
  on consciousness, 296
  publications, 88*n*7, 223, 258, 260*n*7, 299*n*44
  on qualia externalism, 291
  use of "indicator," 260*n*6

Earth and Twin Earth thought-experiment, 242–244, 248
Egan, Frances, 258
Eliminativism, 264–265
Emergentism, 45, 97, 200, 305
Emotion, 16
  belief-desire explanations and, 224*n*3
  consciousness of, 272
  qualitative aspects of, 274–275, 276–277
Enç, Berent, 89*n*15
Entailment
  metaphysical, 89*n*16
  pain-behavior, 76–78
Epiphenomenalism, 96–97
  anomalous monism as form of, 207–209
  causal argument for psychoneural identity and, 114
  consciousness and, 224*n*6
  exclusion argument and, 217
  mental causation and, 199–202
  methodological epiphenomenalism, 320
  physical reductionism and, 327–328
  radical, 219
  supervenience argument and, 217–220
  type epiphenomenalism, 225*n*13
Epistemic (doxastic) sense of appear, 293–295

Epistemological argument against psychoneural identity theory, 115–119
Epistemological criteria for mentality, 18–22
Equipotentiality, 166*n*9
Essence, real *vs.* nominal, 134, 166*n*13
Essential natures, 38–40
Ethics, supervenience theses in, 10
Event, 7
Evolution
of phenomenal consciousness, 322
teleological approach and, 240–241
Spandrel effect in, 322, 336*n*26
Exclusion argument, 127*n*16, 216–217, 219, 223–224
Exclusion principle, 216
Existential quantifier, 171
Experience
first-person point of view and, 282–283
transparency of, 289–295
and phenomenal properties, 277
Experimentation, 198
Explanandum, 103
consciousness as, 317
Explanans, consciousness as, 317
Explanation, as derivation, 106, 107–108
Explanatory arguments, for psychoneural identity theory, 102–110
Explanatory gap, 302–303, 305–306, 333, 334*n*3
closing, 311–315
functional analysis (reduction) and, 315–317, 327
identity reduction and, 313–315
Externalism about qualia, 290–295
Extrinsic mental state, 221–223

Facts, 7
Feelings, 16
Feigl, Herbert, 101, 125, 126*n*7
Filler functionalism, 191*n*14
First-order property, 135, 183
First-person authority, 21–22, 306
subjectivity and, 280–281
See also Direct knowledge of own mental states
First-person point of view, experience and, 282–283
First-person privilege, 19, 21–22
Flanagan, Owen, 296, 297*n*12, 297*n*13, 297*n*19, 298*n*23, 298*n*26, 336*n*31
Fodor, Jerry A.
on mental causation, 201–202, 332
on mind-body supervenience, 335*n*17
publications, 125, 126, 189, 191*n*12, 191*n*13, 224*n*8, 258, 260*n*7, 260*n*10, 261*n*18, 335*n*19, 337*n*41
on reductionism, 311–312
Folk dualism, 31–32
Folk psychology
to anchor psychological concepts, 173, 175–176
See also Commonsense psychology
Formality, of psychological properties, 132
Forrest, Peter, 125
Foster, John, 45–46, 58, 58*n*1, 59*n*14, 60*n*25, 125
Freudian depth psychology, 275
Functional analysis, of pain, 315–317, 327
Functional concept, 133, 280, 304, 318, 330
Functional definition, 134–135, 170, 172–173, 175–177, 180, 316–318, 329
Functionalism, 59*n*15, 122
analytical functionalism, 173
behaviorism and, 136–138
characterization of, 134
criticism of, 189

Functionalism *(continued)*
as philosophy of cognitive science, 129
as physicalism, 177–179
physicalist functionalism, 178
psycho-functionalism, 173
qualia argument against, 179–181
realizer functionalism, 187–188, 191*n*14
role functionalism, 187–188, 191*n*14
See also Causal-theoretical functionalism; Machine functionalism
Functional property, 134–135, 183–186
qualia and, 308–309, 335*n*14
Functional reduction, 328–329, 333
of intentional-cognitive states, 337*n*39
qualia and, 331–332
Functional specification theory, 191*n*14
Functional-state identity theory, 191*n*14
Function-*versus*-mechanism dichotomy, 132, 166*n*8

Garber, Daniel, 58, 60*n*18
Gendler, Tamar Szabo, 59*n*7, 127*n*21, 335*n*15
Genuine/intrinsic intentionality, 26
Gibbons, John, 224
Ginet, Carl, 224*n*4
Global supervenience, 10
Global workplace theory, 278–279
God
creation of C-fiber stimulation and pain, 307
mind-body relation and, 95–96
mind-body union and, 56
as only true substance, 58–59*n*4
teleological approach and, 241
Goldbach's conjecture, 39
Goldman, Alvin I., 28, 190*n*5, 289, 297*n*16, 299*n*39, 299*n*41
Gopnik, Alison, 29*n*11, 298*n*30
Gordon, Robert M, 190*n*5
Gould, Stephen Jay, 336*n*26
Graham, George, 88
Greeting, 75
Grim, Patrick, 336*n*29
Güzeldere, Güven, 296, 297*n*12, 297*n*13, 297*n*19, 298*n*23, 298*n*26, 336*n*31

Habits/propensities, 17
Hardin, C. L., 191*n*10, 335*n*12
Hard problem of consciousness, 303–306
Harman, Gilbert, 126*n*8, 299*n*38, 299*n*43
Harnish, Robert M., 28
Hart, W. D., 58
Hasker, William, 58
Hawthorne, John, 59*n*7, 127*n*21, 335*n*15
Heckmann, Heinz-Dieter, 223
Heil, John, 28, 125, 126*n*5, 165, 167*n*27, 223, 259, 296*n*10, 336*n*30
Hempel, Carl G.
behavioral translation of "Paul has a toothache," 70–71
logical behaviorism of, 68
publications, 87, 88*n*9, 88*n*11, 89*n*23, 335*n*20
Higher-order perception (HOP) theory of consciousness, 283–289, 296, 332
Higher-order thought (HOT) theory of consciousness, 284, 296, 332
Hill, Christopher S.
bodily disturbance theory of pain, 318–319
on consciousness, 296, 297*n*17, 297*n*20, 298*n*34
on first-person point of view, 283
on multiple realizability, 337*n*40
on pain experiences, 318, 319
on phenomenal consciousness, 274
publications, 125, 126*n*9, 189, 337*n*40
somatosensory theory, 336*n*24

Holistic conception of mentality, functionalism's, 138
HOP. *See* Higher-order perception (HOP) theory of consciousness
Horgan, Terence, 223, 334*n*3, 335*n*17
HOT. *See* Higher-order thought (HOT) theory of consciousness
Howell, Robert, 295
Human consciousness, 199–200
Hume, David, 89*n*12
Huxley, Thomas H.
  on animal consciousness, 199
  on consciousness, 265–266, 303
  epiphenomenalism and, 96, 105, 199, 224*n*6
  publications, 126*n*2, 224*n*5, 296*n*7, 334*n*6
Hypothesis testing, 106

Idealism, 33
Identity(ies)
  correlation as, 94
  necessity of, 41–42
  logical rules governing, 107
Identity physicalism, decline of, 129
Identity reduction
  of consciousness, 313–315, 317
  of qualia, 332, 337*n*39, 337*n*40
  and explanatory gap, 313–315
Imitation game, 157. 159
Immaterial, mentality as nonspatial and, 23
Immaterial minds, 39–46, 220. *See also* Substance dualism
Impenetrability
  of matter, 53, 55
  of minds, 53, 55–56
Indeterminancy of interpretation, 232–234
Individuation, principle of, 132–133
Inductive inference, 64
Inductive rule of inference, 102–104, 106–107
Infallibility, knowledge of own mental state and, 19–22
Inference, inductive rule of, 102–104, 106–107
Inference to the best explanation, 102–107, 110, 126, 332
Informational semantics, 235–239
Information processing, cognition as, 161–162
Inputs/outputs
  in causal-theoretical functionalism, 169
  Turing machine, 146–147, 148, 152–154, 156, 158–159
Instrumentalism, about scientific theory, 149–150
Intelligence, Turing test and, 157–158
Intentional inexistence, 23–24
Intentionality
  as-if/derivative, 26
  content, 24–25
  as criterion of the mental, 23–27
  genuine/intrinsic, 26
  interpretation and, 233–234
  linguistic *vs.* mental, 29–30*n*16
  referential, 24–25
Intentional property, 37–38
Intentional states, 15, 27, 29*n*8, 227–228
  causal tasks of, 330
  consciousness and capacity for, 288
  phenomenal knowledge of, 276–277
  reduction of, 328–329, 337*n*39
  supervenience on physical-behavioral properties, 329–330
  See also Internal states; Mental states
Internal states
  functionalist *vs.* behaviorist perspective on, 136–137
  reality of, 150
  of Turing machine, 139, 145–146, 147–148, 152
  See also Intentional states; Mental states

Interpretation, indeterminancy of, 232–234
Interpretation theory, 228–235
  radical interpretation, 229
  See also Charity principle
Intrinsic intentionality, 26
Introspective data, 83–84
Inverted spectrum, 308, 335*n*12
Involuntary memory, 193–194
Itch, 4, 15, 21, 65, 76, 97, 104–105, 107–108, 115–116, 118, 169, 246–247, 268, 302, 305, 307, 332
Itch box, 181–182, 319, 331

Jackson, Frank, 28*n*2, 295, 323, 334, 336*n*27, 336*n*30, 337*n*34
Jacob, Pierre, 223, 258
James, William
  on behavior and mentality, 62
  on consciousness, 266
  on explaining mind-body associations, 105
  on psychoneural correlations, 303, 313
  publications, 88*n*1, 296*n*8, 297*n*13, 334*n*5
  on scope of psychology, 61
Jeffrey, Richard C., 166*n*17
Job description, 131, 329–331
  See also Functional concept
Jolley, Nicholas, 166*n*13

Kim, Jaegwon, 29, 125, 127*n*23, 189, 191, 223, 225, 226, 334, 337
Kind, Amy, 296
Knowledge
  mental causation and, 197–198
  physical, 325
  propositional, 324–325
  subjective *vs.* objective, 37
  third-person, 19, 281
  See also Direct knowledge of own mental states; First-person authority
Knowledge argument, 323–326, 334, 337*n*34
Koons, Robert C., 125, 334
Kripke, Saul, 80, 90*n*27, 119, 125, 127*n*20, 307, 335*n*9

Language
  content of belief and, 244–248, 250
  intentionality and, 25–26, 29–30*n*16
  mentalistic, 87, 100
  physical, 100
  psychological, 65–66
Laplace, Pierre de, 99
Lashley, Karl, 166*n*9
Latham, Noa, 58
Leibniz, Gottfried, 34
  causation and, 54
  Leibniz's mill, 44
  on consciousness and material things, 44–45, 59*n*12
  on mind-body relation, 95
LePore, Ernest, 223
Levin, Janet, 295
Levine, Joseph, 260*n*7, 333, 334*n*3
Lewis, David
  analysis of causation, 225*n*14
  publications, 189, 190*n*2, 190*n*3, 225*n*14, 225*n*17, 259*n*1, 261*n*17, 336*n*32
  Ramsey-Lewis method, 170–172
Lewontin, Richard, 336*n*26
Linguistic communities, content of belief and, 244–248, 250
List, Christian, 224
Loar, Brian, 127*n*19, 336*n*31
Locke, John, 59*n*10, 166*n*13
  the prince and the cobbler, 42–43
Loewer, Barry, 223, 258, 260*n*7
Logical behaviorism (philosophical/analytical behaviorism), 68–69
  ontological behaviorism and, 78–79
  See also Behaviorism

Logical positivism, 69
Lowe, E. J., 58, 125, 223
Ludlow, Peter, 28, 259, 334
Lycan, William G., 28, 165, 166*n*8, 296

Macdonald, Cynthia, 224, 259
Macdonald, Graham, 224
Machine functionalism, 129–167
  claims and motivations, 147–151
  computationalism, 160–165
  functionalism and behaviorism, 136–138
  functional properties and their realizers, 134–135
  further issues for, 151–156
  multiple realizability and functional conception of mind, 130–134
  See also Turing machines, Turing test
Malcolm, Norman, 88
Malebranche, Nicolas, 34, 96
Mark of the mental, 17–27
Martin, Norah, 259
Materialism, 11–14, 32–33, 301
  causal argument for, 49
  defined, 11
  See also Physicalism
Material monism, 32
Material things, as incongruous with mental states, 39–46
Matthews, Robert, 260*n*14
McLaughlin, Brian, 28, 28*n*1, 126*n*10, 127*n*19, 189, 191*n*14, 223, 225*n*13, 260*n*7
Meaning
  belief, behavior, and, 229–235
  supervenience on internal physical-psychological states and, 244
  verifiability criterion of, 69
Mele, Alfred, 223
Melnyk, Andrew, 125, 126*n*10, 127*n*23
Melzack, Ronald, 166*n*7
Memory, 304
  involuntary, 193–194
Mendola, Joseph, 258
Mental
  mark of, 17–27
  requirement of rationality and coherence for, 205
Mental acts, 67
Mental-behavior entailments, defeasibility of, 74–78
Mental causal efficacy, nonreductive physicalism and, 123–125
Mental causation, 46–50, 193–226
  agency and, 195–197
  anomalous monism as form of epiphenomenalism, 207–209
  argument for psychoneural identity theory, 110–114
  counterfactual account of, 209–213
  exclusion argument, 214–217, 219, 223–224
  exclusion principle, 216–217, 224
  extrinsicness of mental states, 220–223
  loss of, 332
  mental realism, epiphenomenalism, and, 197–202
  physical causal closure and, 214–217
  problem of, 4–5
  psychophysical laws and anomalous monism, 202–207
  supervenience argument, 217–220
Mental content, 227–261
  causal-correlational approach, 235–239
  causal-explanatory efficacy of wide content, 254–256
  content externalism, 241–248, 254–257
  informational semantics, 235–239
  interpretation theory, 228–235
  metaphysics of wide content states, 248–251
  misrepresentation and the teleological approach, 239–241

Mental content *(continued)*
narrow content, 241–246, 251–254
possibility of narrow content, 251–254
wide content, 241–248, 235–239, 256–257
wide content and self-knowledge, 256–257
Mentalism, 82
Mentality
behavior and, 61–62
conception of, 27
consciousness and, 263
intentionality and, 23–27
as nonspatial, 22–23
physicalist view of, 221
physical reduction of, 327–328
properties of, 6–7
relation to physicality, 203
Mental kind, 133–134
functional definitions of, 170–172
Mental language, behavior and, 87
Mental phenomenon, 3
conscious states, 27
epistemological criteria, 18–22
intentionality as a criterion of the mental, 23–26
intentional states, 27
mentality as nonspatial, 22–23
varieties of, 14–17
Mental property
causal efficacy of, 327, 337*n*35
causal powers of, 184–186
mental causation and, 208–209
multiple realizability of, 129–134, 183–186
relation to physical property, 9, 12–13
Mental property epiphenomenalism, 208
See also Epiphenomenalism
Mental realism, 137–138, 197–202
Mental state, 3
behavior as evidence for attribution of, 87
in causal-theoretical functionalism, 169–170
as conscious state, 272
extrinsicness of, 220–223
folk psychology and attribution of, 173, 175–176
functional analysis of, 327
functionalist *vs.* behaviorist perspective on, 136–138
incongruity with material things, 39–46
knowledge of own mental states (*see* Direct knowledge of own mental states)
ontological behaviorism and, 78
relation to behavior, 74
subconscious, 36–37
unconscious, 36–37
varieties of, 14–17
See also Intentional states; Internal states; Multiple realization of mental states
Mental-to-mental causation, 194, 198
epiphenomenalism and, 201
mind-body supervenience and, 219
Mental-to-physical causation, 194, 197, 198
epiphenomenalism and, 201, 202
mind-body supervenience and, 219
physical closure principle and, 215–217
Menzies, Peter, 224
Metaphysical entailment, 89*n*16
Metaphysical possibility, 120, 337*n*37
See also Possibility, real, Conceivability
Metaphysical thesis of physicalism, 324, 326
Meter, concept of, 80–81, 90*n*28
Methodological behaviorism, 83, 84–85

Millikan, Ruth G., 165, 258
Mind
  behaviorism and Cartesian conception of, 63–66
  brain and, 6, 87
  categorizing things as having/not having, 1–3, 57
  as computing machine (*see* Machine functionalism)
  functional conception of, 130–134
  philosophy of, 2–5
  as thinking substance, 35–39
  See also Substance dualism
Mind-body causal interaction, 34–35, 46–50
Mind-body dependence, 12–13
Mind-body dualism, 5, 31–32
  closure principle and, 214
  mind-body supervenience and, 12
  See also Property dualism, Substance dualism
Mind-body problem, 3–5
  consciousness and, 301–302
Mind-body relation, causal approaches to, 95–98
Mind-body supervenience, 8–11, 166*n*6, 334*n*3, 335*n*17
  epiphenomenalism and, 217–220
  explanatory gap and, 302–303, 305–306
  nonreductive physicalism and, 123, 124
  physicalism and, 12–14
  See also Supervenience
Mind-body union, Descartes and, 47–50, 56
Mind-brain correlations, 91–93
  making sense of, 93–98
  mind-brain correlation thesis, 92
Mind-brain identity theory. *See* Psychoneural identity theory
Mind reading, 270, 297*n*16
Mind-world relation, intentionality and, 24
  See also word-to-world (language-to-world) relation
Minimal physicalism, 13, 166*n*6
Misrepresentation, 239–241, 258
Mnemic causation, 85
Modal argument against psychoneural identity theory, 119–120
Mode of presentation, knowledge and, 118–119
Monism, 32
  anomalous, 207–209, 223
  material, 32
  neutral, 96
Moods, 275, 278
  and consciousness representationalism, 293
Morgan, C. Lloyd, 305
Motions and noises, 67, 71, 72, 75, 76
Mousetrap, as functional concept, 135, 166*n*14
Multiple draft theory, 279
Multiple realizability (realization) of mental properties (states), 89*n*19, 121–122, 125,129–134, 183–187, 188, 189, 312, 328, 337*n*40
  functionalism and, 176–177
Mundale, Jennifer, 125
Murdoch, Dugald, 58
Mystery of consciousness, 266, 305–306

Nagel, Ernest, 225*n*16
Nagel, Thomas
  on bat consciousness, 267–271
  on consciousness, 266, 288
  on consciousness and mind-body problem, 301
  on first-person point of view, 282, 297*n*15
  on single subject for each experience, 298*n*33

Narrow content, 241–248, 258
  possibility of, 251–254
Natural selection
  consciousness and, 322
  teleological approach and, 240–241
Neander, Karen, 258
Necessity of identities (NI), 41–42, 59*n*9
Nemirow, Lawrence, 336*n*32
Neural-physical domain, causal-explanatory closure of, 320
Neural-physical properties, 222
Neuroscience, consciousness and, 317–323
Neutral monism, 96
Ney, Alyssa, 29*n*3
NI. *See* Necessity of identities (NI)
Nida-Rümelin, Martine, 191*n*10, 299*n*46
Nisbett, Richard, 29*n*11, 298*n*30
Nomic-derivational approach, to counterfactuals, 210
Nominal essence, 134, 166*n*13
Nomological account of causation, 206, 209
Nomological danglers, 101, 104
Nonconceptual content, 285–286
Nonreductive physicalism, 11, 13–14, 29*n*5, 56–58, 122–126
Non-rigid designators, 119
Nonspatiality, of mentality, 22–23
Nuccetelli, Susan, 259
Null extension, 174
Null set, 190*n*5

Objects, relation to events and states, 7
Observational beliefs, 237–238
Occam's (Ockham's) razor, 98–99
Occasionalism, 96
Occasion sentences, 259*n*3
Occurrence, 7
O'Connor, Timothy, 28
Olson, Eric T., 59*n*8
Ontological behaviorism, 78–80
Ontological physicalism, 29*n*4, 166*n*6
Ontological primacy of the physical in relation to the mental, 12–13
Ontological scheme, 7
Ontology, 6
Overdetermination, 216

Pain
  bodily disturbance theory of, 299*n*47, 318–319, 336*n*24
  Cartesian conception of the mind and, 65
  as causal intermediary, 89*n*22
  causal-nomological relations of, 172
  causal powers of, 337*n*35
  C-fiber stimulation and, 99–101, 105, 106, 107–110, 119–120, 307–308, 312–313, 326–327
  commonality to all instances of pain, 179–180
  congenital incapacity for, 336*n*25
  cross-wired brain and, 181–182
  functional analysis of, 315–317, 318, 327
  function of, 331
  in higher-order theories of consciousness, 285–286
  knowledge of own, 20
  meaning of, 25, 97
  mind-body supervenience and, 302–303
  ontological behaviorism and, 78–79
  pain behavior and, 76–78, 80–82
  phenomenal aspect of, 304–305
  psychoneural identity theory and, 99–101, 105, 106, 107–110
  Ramsey-Lewis method and, 170–172
  realizer functionalism on, 187–188
  reductive explanation of, 312–313
  representation of, 318–319
  role functionalism on, 187–188
  somatosensory theory of, 336*n*24
  as tissue-damage detector, 131, 133–134, 166*n*12

Pain box, 331
Pairing problem, for substance dualism, 50–54, 55–56, 58, 127*n*15
Papineau, David, 127*n*14, 223, 226*n*20, 334
Pauen, Michael, 223
Pavlov, Ivan, 264, 296*n*3, 323
Perception, 197–198
Perry, John, 334
Phases of moon, tides and, 95, 96–97
Phenomenal concept strategy, 117
Phenomenal consciousness, 15, 272–277
  access consciousness and, 279–280
  causal status of, 319–321
  evolution of, 322
  explanatory gap and, 302–303
  hard problem and, 303–306
  higher-order thought theory and, 286–287
  qualia representationalism and, 333
  scientific investigation of, 321–322
  supervenience on physical properties, 306–311
Phenomenal properties, 273. *See also* Qualia, Phenomenal consciousness
Philosophical behaviorism, 62, 68
Philosophy of mind, defined, 2–5
Physical causal closure. *See* Causal closure of physical domain
Physical events, 7
  relation to mental events, 4
Physicalism, 11–14, 29*n*3, 32–33, 301–302
  antiphysicalist argument, 323–326
  causal argument for, 54
  defense of, 334
  defined, 11
  functionalism as, 177–179
  Jackson's physicalism, 323
  limits of, 326–333
  metaphysical thesis of, 324, 326
  mind-body supervenience and, 12–14
  minimal physicalism, 13
  nonreductive physicalism, 13, 29*n*5, 122–126
  ontological physicalism, 29*n*4
  qualia supervenience and, 310–311
  realization physicalism, 123, 124, 130
  reductionist physicalism, 219–220
  reductive (type) physicalism, 14, 57, 119, 122–125, 129
  substance physicalism, 13, 57
  supervenience physicalism, 123, 124
  token physicalism, 14, 29*n*7, 122, 207
Physical law, causation and, 205–207
Physical property
  relation to mental properties, 9, 12–13
  supervenience of qualia on, 335*n*14
  supervenience of consciousness on, 306–311
Physical realizationism, 127*n*23
Physical realizer, realization, 130, 135, 146, 152, 177–178, 189, 203
  of Turing machines, 144–147
  See also Realizer, Realization
Physical requirement
  on causal-theoretical functionalism, 178
  on machine functionalism, 178
Pineal gland, as seat of the soul, 47, 55, 95
Place, U. T., 126*n*4
Plantinga, Alvin, 45, 58, 59*n*13
  on Leibniz's mill, 45
Plato, 31
Point of view
  consciousness and, 269–270
  first-person, 282–283
Poland, Jeffrey, 125
Polger, Thomas W., 125
Positivism, logical behaviorism and, 68–69
Possibility, real, 39, 120, 309, 325, 335*n*15
  conceivability and, 39, 120, 335*n*15

Possible-world semantics, of counterfactuals, 210–212, 225*n*15
Predicate constants, 171
Predicate variables, 171
Preestablished harmony between mind and body, 95
Prince and the cobbler, the, 42–43
Princess Elisabeth of Bohemia, 47–50, 60*n*20, 60*n*21
Principle of charity, 225*n*11, 230–232, 259*n*3
Principle of individuation, 53, 132–133
Principle of inference to the best explanation, 102–104, 106–107
Principles of rationality, 204–206
Prinz, Jesse, 28
Privacy of knowledge, of own mental states, 19, 27, 65
Probabilistic automaton, 143–144, 147
Process, 7
Proper name, 119
Property dualism, 13–14, 29*n*5, 56–58, 124
Property, 7
  functional, 134–135
  functional reduction of, 328–329
Propositional attitude, 15, 204, 227
  emotions and, 16
  knowledge of own, 20
Propositional content, 328
Propositional knowledge, 324–325
Proprioception, 19
Proust, Marcel, 193–194, 224*n*1
Psycho-functionalism, 173
Psychological eliminativism, 78
Psychological expressions, behavioral definability of, 68–69
Psychological reality, 150, 167*n*28, 177–179
Psychology
  behaviorism and, 61, 82–86
  commonsense (*see* Commonsense psychology; Folk psychology)
  nonreductive physicalism and autonomy of, 124–125
  scientific, 173, 175–176, 177, 178–179
Psychoneural correlations
  explanations of, 126*n*12
  hard problem of consciousness and, 303–305
  psychoneural identity theory and, 101–105, 109–110
Psychoneural identity theory, 91–127, 326
  arguments against, 114–122
  causal argument for, 110–114
  cross-wired brain problem and, 182
  epiphenomenalism and, 217
  explanatory arguments for, 102–110
  psychophysical laws and, 203
  reductive and nonreductive physicalism, 122–125
  simplicity argument for, 98–102
Psychophysical anomalism, 206, 209
Psychophysical counterfactuals, 212–213
Psychophysical laws, 202–207, 224–225*n*9, 225*n*10
  evaluation of psychophysical counterfactuals and, 212–213
Putnam, Hilary
  on content, 258
  as critic of functionalism, 189
  functionalism and, 139
  multiple realization argument and, 121–122, 337*n*40
  publications, 88, 89*n*17, 125, 129, 165, 167*n*19, 260*n*11, 260*n*13
  Twin Earth thought-experiment, 242–244

Qualia, 273–277, 297*n*18
  definitions of, 15, 295, 317
  differences and similarities of, 333
  explanatory role in brain and behavioral sciences, 319–321

functional definition of, 317
functional properties and, 308–309
functional reduction and, 331–332
identity reduction of, 332
nature of consciousness and, 268
methodological qualia epiphenomenalism, 320
qualia epiphenomenalism, 334
qualia externalism, 290–295
qualia representationalism, 289–295, 332–333
supervenience of, 306–311, 335*n*14
as properties of brain states, 294
scientific theory of, 321–322
Qualia inversion, 179–181, 182, 191*n*10, 308–309, 331
Qualia nihilism, 264–265
Quine, W. V., 259*n*1, 334*n*1

Radical behaviorism, 82
Radical epiphenomenalism, 219
Radical interpretation, 229
See also Interpretation theory
Radical translation, 259*n*1
Ramseification, 170–171. *See also* Ramsey–Lewis method
Ramsey, Frank P., 170
Ramsey-Lewis method, 166*n*16, 170–172
psychology underlying, 172–177
Rasmussen, Joshua, 58
Rational action, 73
Rationality
principles of, 204–206
requirement of, 259*n*3
Rationalizations, of actions, 196–197
Raw feels, 15, 273
Real essence, vs. nominal essence, 134, 166*n*13
Realism, about scientific theories, 150–151
Reality, psychological, 177–179
Realization, 130, 135, 144–147, 152, 177–178, 189, 203
See also Multiple realizability
Realization physicalism, 123, 124, 127*n*23, 130
Realizer functionalism, 187–188, 191*n*14
Realizer
of functional properties, 135
physical, of Turing machines, 144–147
Recognitional concepts, 117, 324
Reduction, 305, 311–315
functional reduction, 315–317
identity reduction, 313–315
of mentality, 327–328
See also Functional reduction
Reductionism, reductionist physicalism, 219–220
See also psychoneural identity theory, reduction
Reductive explanation, 305, 311–315
functional analysis and, 315–317
Reductive physicalism (type physicalism), 14, 57, 119, 122–125, 129
psychoneural identity theory as form of, 122
Referential intentionality, 24–25
Relational properties, 221, 249
pairing problem and, 51–52
Relations, 7
Representation, 24, 25, 26, 160, 164, 221, 228, 239–241, 268, 278, 293, 294, 319
mental representation, 25, 85
satisfaction conditions of, 278, 293
misrepresentation, 239–241, 259
representational vehicle, 294
Representational content, 24–25, 164, 220, 221, 222, 228, 241, 249–250, 251, 255, 278–280, 290, 292
Representationalism
about consciousness, 27, 280, 289–295
qualia, 289–295, 318–319

Representational properties, beliefs and, 249
Resemblance, causal-nomological, 186
Rey, Georges, 88, 258, 260*n*7, 265, 296*n*6, 335*n*16
Rigid designators, 59*n*9, 119
Rimbaud, Arthur, 308, 335*n*11
Robb, David, 28
Rock objection, to higher-order theories of consciousness, 289
Role functionalism, 187–188, 191*n*14
Rosenthal, David, 284, 296, 299*n*37
Ross, Don, 188, 189, 191*n*14, 191*n*15
Rowlands, Mark, 299*n*39
Rozemond, Marleen, 58, 59*n*11
Ryle, Gilbert, 87, 89*n*23, 89*n*24

Satisfaction conditions, of representations, 278, 293
Scanner-printer, of Turing machines, 139, 145
Schank, Roger, 161
Schneider, Susan, 295
Science
  behaviorism in, 82–86
  causal closure and, 215
  explaining correlations in, 105–106
  metaphysics of scientific theories, 149–150
  study of consciousness, 265–267, 317–323
Scientific psychology, as underlying theory to be Ramseified, 173, 175–176, 177, 178–179
Searle, John R.
  on causal power of brain, 167*n*30
  Chinese room argument, 161–164
  publications, 29*n*15, 30*n*17, 165, 167*n*27, 167*n*29, 189, 298*n*28
  on strong AI, 161
Second-order perception. *See* Higher-order perception (HOP) theory of consciousness
Second-order properties, 135, 183–184
Segal, Gabriel, 258, 261*n*18
Self, consciousness and notion of, 287–288
Self-awareness, 306
Self-interpretation, 234–235
Self-knowledge, wide content and, 256–257, 259
Sellars, Wilfrid, 29*n*16, 265, 322
Semantic knowledge, 163
Sensations
  mental phenomena involving, 15
  qualitative features of, 180
Sentences, content, 15
Shaffer, Jerome, 127*n*17
Shapiro, Lawrence, 125, 166*n*14
Shepard, Roger, 189
Shoemaker, Sydney, 189, 191*n*10, 299*n*46, 335*n*12
Siewart, Charles, 296
Simplicity argument, for psychoneural identity theory, 98–102
Simulation theory, 190*n*5, 297*n*16
Skinner, B. F., 82, 84, 88, 90*n*29, 166*n*15
Smart, J. J. C.
  psychoneural identity theory and, 98, 101, 104, 106, 109, 116, 125, 332
  publications, 125, 126*n*4, 126*n*5, 126*n*11, 127*n*17, 127*n*18
Smith, Barry C., 259
Smith, Michael, 224*n*3
Socrates, 97
Solipsism, Nagel's discussion of consciousness and, 269–270
Solubility, realist *vs.* instrumentalist analysis of, 136–137
Somatosensory theory of pain, 336*n*24
Sosa, Ernest, 223, 260*n*14
Soul
  Cartesian, 47, 52, 55, 56, 60*n*23
  Platonic, 31–32
  See also Substance dualism

Space
  immaterial minds in, 54–56
  physical causation and, 53
Spandrel effect, 322, 336*n*26
Spatial relations, pairing problem and, 51
Spectrum inversion, 293, 299*n*46, 331
Speech, content of belief and, 245–248, 250
Spinoza, Baruch, 34, 96
Spurrett, David, 188, 189, 191*n*14, 191*n*15
Stalnaker, Robert, 109, 126*n*12, 260*n*6, 260*n*7, 314, 334, 336*n*21
Stampe, Dennis, 260*n*7
State consciousness, 272
Staudacher, Alexander, 223
Stich, Stephen, 80, 90*n*26
Stimulus generalization, 189
Stimulus-response-reinforcement model, 84–85
Stoljar, Daniel, 125, 334
Stoothoff, Robert, 58
Stoutland, Frederick, 224*n*3, 224*n*4, 225*n*13
Strawson, Galen, 30*n*17
Strict laws, 213, 225*n*12
Strong AI, 161
Stubenberg, Leopold, 296
Subconscious mental states, 36–37
Subconsciousness, 20–21
Subject consciousness, 272
Subjectivity
  consciousness and, 280–283
  first-person authority and, 280–281
Substance
  defined, 7, 33
  mental *vs.* material, 33–34
Substance dualism, 29*n*5, 32–35
  arguments for, 35–46
  immaterial minds in space, 54–56
  mental causation in, 46–50, 202
  pairing problem and, 50–54, 55–56
  property dualism and, 56–58
Substance physicalism, 13, 57
  nonreductive physicalism and, 123, 124
Substitution rule, of identity, 107
Super blindsighter, 279–280
Super-Spartans, 76–77
Supervenience
  Burge's thought-experiment and, 245
  global, 10
  of beliefs, 253
  of mentality on brain states, 93
  Putnam's thought-experiment and, 244
  qualia, 306–311, 335*n*14
  strong, 9
  See also Mind-body supervenience
Supervenience argument, 217–220, 223–224
Supervenience bases, multiple, 334*n*2
Supervenience physicalism, 123, 124
Surfaces, 33–34
Swinburne, Richard, 58
Synesthesia, 335*n*11
Syntax, 163–164, 221
  as incapable of generating meaning, *See* Chinese room argument

Tape, of Turing machines, 139, 143, 144
Telekinesis, 111
Teleological approach, to content, 239–241
Theory
  behaviorism and objective testability of, 82–86
  metaphysics of scientific, 149–150
  simplicity in constructing, 99
Theory theory, of commonsense psychology, 190*n*5
Thought-experiments
  arthritis and tharthritis, 244–248

Thought-experiments *(continued)*
Earth and Twin Earth, 242–244, 248
Mary, a vision scientist, 323–326, 336–337*n*33
Token physicalism, 14, 29*n*7, 122, 207
Topic-neutral translation, of phenomenal reports, 116
Translatability thesis, of behaviorism, 68–71, 89*n*23
Transparency, 29*n*10
of experience, 289–295
of mind, 20–22
"Transporter," *Star Trek,* 8, 11
Trope theory, 224
Truth, of beliefs, 230–232
Truth conditions, for content individuation, 242
Turing, Alan M., 139, 156, 165, 167*n*22
Turing machines, 139–144
inputs/outputs, 146–147, 148, 152–154, 156, 158–159
internal states, 145–146, 147–148, 152
machine tables, of Turing machines, 140–142
mathematical theory of computability and, 139, 166*n*17
physical realizers of, 144–147
psychology represented by, 147
universal Turing machines, 143
Turing's thesis, 158–159
Turing test, 156–159, 165, 167*n*24
Tye, Michael, 296, 299*n*43, 335*n*12
Type epiphenomenalism, 225*n*13
Type physicalism (reductive physicalism), 14, 57, 119, 122–125, 129

Unary notation, 140
Unconscious mental states, 36–37
Unity of consciousness, 285

Van Fraassen, Bas, 126*n*8, 167*n*20
Van Gulick, Robert, 295
Van Horn, Luke, 58
Variable realization, 121, 127*n*22
Veillet, Benedicte, 127*n*19
Velmans, Max, 295
Verbal behavior, 70–72
Verbal reports of inner experience, 83–84
Verifiability criterion of meaning, 69
Vernacular psychology, 176, 190*n*8
See also Commonsense psychology
Volitions, 16
von Eckardt, Barbara, 28

Walter, Sven, 28, 127*n*19, 223, 334
Watson, J. B., 61, 88*n*3
Weiskrantz, Lawrence, 298*n*27
White, Stephen, 261*n*18
Wide content, 241–248, 258, 261*n*16
causal-explanatory efficacy of, 254–256
metaphysics of, 248–251
self-knowledge and, 256–257, 259
William of Ockham, 98, 126*n*6
Wilson, Timothy DeCamp, 29*n*11, 298*n*30
Witmer, Gene, 190
Wittgenstein, Ludwig, 24, 29*n*13, 64–65, 88*n*6
Word-to-world (language-to-world) relation, meaning and, 163–164
Wright, Crispin, 259
Wright, Edmond, 334

Yablo, Stephen, 223

Zimmerman, Dean, 58
Zombies, 120, 298*n*26, 309, 329, 331
Zombie worlds, 307